Grades 5, 6, and 7
New Jersey Assessment of Skills and Knowledge

TECHNICAL REPORT

DECEMBER 2006

PTM #1506.09
Copyright © 2006 by New Jersey Department of Education
All rights reserved

NJ ASK 2006 GRADES 5–7 TECHNICAL REPORT
TABLE OF CONTENTS

PART 1: INTRODUCTION

1.1 Description of New Jersey's Overall Assessment System

New Jersey's Office of Evaluation and Assessment (OEA) coordinates the development and implementation of New Jersey's state-required assessment program designed to measure the extent to which all students at the elementary-, middle-, and secondary-school levels have attained New Jersey's Core Curriculum Content Standards (CCCS). New Jersey's state required assessment system includes the following components:

Elementary School:

- Grade 3 New Jersey Assessment of Skills and Knowledge
- Grade 4 New Jersey Assessment of Skills and Knowledge

Middle School:

- Grade 5 New Jersey Assessment of Skills and Knowledge
- Grade 6 New Jersey Assessment of Skills and Knowledge
- Grade 7 New Jersey Assessment of Skills and Knowledge
- Grade Eight Proficiency Assessment (GEPA)

High School:

- High School Proficiency Assessment (HSPA)

In addition, the statewide assessment program currently includes two tests for special populations:

- Alternate Proficiency Assessment (APA), for students with severe disabilities
- Special Review Assessment (SRA), for students who have not demonstrated proficiency in one or more content areas of the HSPA

1.2 Brief Summary of the NJ ASK for Grades 5–7

This Technical Report focuses on the 2006 New Jersey Assessment of Skills and Knowledge (NJ ASK) administered as an operational test at grades 5, 6, and 7 in the content areas of Language Arts Literacy and Mathematics in spring 2006. The tests provide an indication of student progress toward achieving the knowledge and skills identified in New Jersey's CCCS, and the tests fulfill the requirements under the federal No Child Left Behind (NCLB) Act.

The NJ ASK is designed to give an early indication of the progress students are making in mastering the knowledge and skills described in the CCCS. The results are to be used by schools and districts to identify strengths and weaknesses in their educational programs. This process is expected to lead to improved instruction and better alignment with the CCCS. The results may also be used, along with other indicators of student progress, to identify those students who may need instructional support in any of the content areas. This support, which could be in the form of individual or programmatic intervention, represents a means to address any identified knowledge or skill gaps.

The NJ ASK Language Arts Literacy and Mathematics scores at grade 5–7 are reported as scale scores. The scores range from 100–199 (Partially Proficient), 200–249 (Proficient), and 250–300 (Advanced Proficient). The scores of students who are included in the Partially Proficient level are considered to be below the state minimum of proficiency and those students may be most in need of instructional support. The standard-setting procedures used for determining proficiency levels are detailed in Part 3 of this Technical Report.

1.3 Spring Testing and Accountability

All testing for NJ ASK is completed in the spring of each year. Recall that the NCLB goal for every school in the state is Proficiency as defined by the New Jersey Board of Education. Testing is conducted in the spring of each year to allow school staff and students the greatest opportunity to achieve the goal of Proficiency.

Data for this Technical Report were collected during the spring administration in April 2006. However, the analyses presented in some Parts (e.g., Part 3 – Standard Setting and Part 7 – Scaling and Equating) of the Report are based on a subset of the total student population. For example, the NJ ASK 2006 grades 5–7 standard setting included student data based on a sample of early returned data (available in May) that consisted of about 14% of the student population. Using a subset of early return data for standard setting was necessary in order to meet reporting timelines established by NCLB. When subsets of data from the total student population are used for analyses, the student N-counts are provided.

Data analyses for the total student population are based on a dataset made available to Riverside Publishing in July. It should be noted that normal reviews of the NJ ASK 2006 grades 5-7 data took place after July and thus there may be very slight differences between the data aggregated in this report and similar data aggregations based on more recent versions of the NJ ASK 2006 dataset.

1.4 NJ ASK Organizational Support

New Jersey's OEA coordinates the development and implementation of the NJ ASK for grades 5–7. In addition to planning, scheduling, and directing all NJ ASK activities, the staff is extensively involved in numerous test reviews, security, and quality-assurance procedures. In November 2005, Riverside Publishing provided an unsolicited proposal to assist the state in developing the NJ ASK for grades 5–7 for an April 2006 test administration. Riverside is the

primary contractor working in partnership with Measurement Incorporated. The major Riverside activities include program management, development of all test materials (test booklets, answer documents, and ancillary materials), and psychometric support, including standard setting. The major Measurement Incorporated activities include enrollment verification; distribution of all materials; receiving, scanning, editing, and scoring the answer documents; scoring open-ended responses; and creating, generating, and distributing all score reports of test results to students, schools, districts, and the state.

1.5 Purpose of the Technical Report

This Technical Report provides information about the technical characteristics of the 2006 administration of the NJ ASK for grades 5, 6, and 7. While some parts of this report are accessible to everyone, its intended audience is experts in psychometrics and educational research. This report is best understood with a working knowledge of measurement concepts such as reliability and validity, and statistical concepts such as correlation and central tendency. For some chapters, the reader is presumed to have basic familiarity with advanced topics in measurement and statistics such as item response theory (IRT).

The report provides extensive detail about the development and operation of NJ ASK. The traditional concerns with a program are often labeled reliability and validity. The empirical reliability and validity of the assessments are reported explicitly in this document. While reliability (Part 12) is relatively straightforward, the steps in creating the program and putting it into operation are all aspects of validity (Part 13). The validity of any assessment stems from the steps taken in planning it, the processes of developing the content of the tests, the processes of consulting with stakeholders, the processes of communicating about the test to users, the processes of scoring and reporting, and the processes of data analysis. Each is an inherent part of validity. In short, while there is a specific chapter devoted to validity, this document provides much but not all of the evidence needed to assess the validity of the NJ ASK.

In reading this technical report, it is critical to remember that the testing program does not exist in a vacuum; it is not just a test. It is one part of a complex network intended to help schools focus their energies on dramatic improvement in student learning. NJ ASK is an integrated program of testing, accountability, and curricular and instructional support. It can only be evaluated properly within this full context.

PART 2: TEST DEVELOPMENT PROCESS

2.1 Introduction

The New Jersey Assessment of Skills and Knowledge (NJ ASK) was first administered in 2006 as an operational test at grades 5, 6, and 7 in the content areas of Language Arts Literacy and Math. The tests provide an indication of student progress toward achieving the knowledge and skills identified in the Core Curriculum Content Standards (CCCS), and the tests fulfill the requirements under the No Child Left Behind (NCLB) Act.

2.2 Test Specifications

Riverside Publishing content experts and the New Jersey Department of Education (NJDOE) developed a directory of test specifications and sample items for each content area. These specifications describe the test, format of the items, and the scores to be generated by the test. This document serves as the foundation for all test item development.

Riverside and the NJDOE rely upon their expertise and the Core Curriculum Content Standards to design a test that is universally accessible to all students in grades 5, 6, and 7 and is composed of test questions that are age- and grade-appropriate. The material in the test specifications is designed for use by curriculum specialists and teachers to improve instructions at the district, school, and classroom levels.

The NJ ASK 2006 administered to students in grades 5, 6, and 7 is designed to measure the same Core Curriculum Content Standards as the NJ ASK administered to students in grades 3 and 4. Brief descriptions of the test content measured in Language Arts Literacy and Mathematics are presented in the following sections.

Language Arts Literacy

The Language Arts Literacy section of each test measures students' achievements in reading and writing. Students read passages selected from published books, newspapers, magazines, and everyday text and respond to related multiple-choice questions, open-ended questions and a writing task.

The Language Arts Literacy assessment currently assesses knowledge and skills in the following clusters (A cluster is a group of related test questions on a single topic.):

- Writing
- Reading
    - Working with Text
    - Analyzing Text

For an in-depth description of the NJ ASK Language Arts Literacy assessment, including specifications, visit the following page on the NJDOE Web site:

http://www.njpep.org/assessment/njask_lal/Overview_njask_lal.pdf

Mathematics

The Mathematics section of each test measures students' ability to solve problems by applying mathematical concepts. The NJ ASK assesses four Core Curriculum Content Standards in Mathematics:

- Number Sense and Numerical Operations
- Geometry and Measurement
- Patterns and Algebra
- Data Analysis, Probability, and Discrete Mathematics

A process cluster, Problem Solving, is also reported on score reports. The process cluster refers to test questions that measure mathematical problem-solving ability. Each test question on the Mathematics assessment measures one content cluster and may contribute to the process cluster. Each content cluster in Mathematics may contain one of the three open-ended items. For an in-depth description of the NJ ASK Mathematics Test Specifications visit the following page on the NJDOE Web site at:

http://www.njpep.org/assessment/TestSpecs/MathNJASK/index.html

Table 2.2.1
2006 New Jersey Assessment of Skills and Knowledge (NJ ASK)
Total Points Possible by Content Area – Grades 5, 6, and 7

Language Arts Literacy	Grade 5	Grade 6	Grade 7
Total	**41 points**	**48 points**	**48 points**
Writing	5 points	12 points	12 points
Writing/Picture	5 points		
Writing/Persuasive Prompt		12 points	12 points
Reading	36 points	36 points	36 points
Working with Text	16 points	15 points	12 points
Analyzing Text	20 points	21 points	24 points
Mathematics	**Grade 5**	**Grade 6**	**Grade 7**
Total	**39 points**	**39 points**	**39 points**
C4.1 - Number Sense & Numerical Operations	10 points	9 points	10 points
C4.2 - Geometry & Measurement	9 points	10 points	9 points
C4.3 - Patterns & Algebra	10 points	10 points	10 points
C4.4 - Data Analysis, Probability, & Discrete Math	10 points	10 points	10 points
Problem Solving	37 points	37 points	37 points

2.3 Test Blueprints

The following tables outline the test construction blueprints. The actual test map for each grade and content area for the 2006 NJ ASK is included.

Table 2.3.1
Test Construction Map for Grade 5 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC[1] (Number of Items)	OE[2] (Number of Items)	WT[3] (Number of Items)	Total Points
Picture Prompt		0	0	1	5
Narrative	1				
AT		4-6	0-2	0	8-12
WT		4-6	0-2	0	8-12
Everyday Text	1				
AT		2-6	0-2	0	8-12
WT		4-8	0-2	0	8-12
Total Items		**20**	**4**	**1**	
Total Points		**20**	**16**	**5**	**41**

[1] MC – Multiple Choice
[2] OE – Open-ended
[3] WT – Writing Task

Table 2.3.2
Actual Test Map for 2006 Grade 5 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC (Number of Items)	OE (Number of Items)	WT (Number of Items)	Total Points
Picture Prompt		0	0	1	5
Narrative	1				
AT		6	1	0	10
WT		4	1	0	8
Everyday Text	1				
AT		2	2	0	10
WT		8	0	0	8
Total Items		**20**	**4**	**1**	
Total Points		**20**	**16**	**5**	**41**

Table 2.3.3
Test Construction Map for Grade 6 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC (Number of Items)	OE (Number of Items)	WT (Number of Items)	Total Points
Persuasive Prompt		0	0	1	12
Narrative	1				
AT		4-6	0-2	0	6-12
WT		4-6	0-2	0	6-12
Everyday Text	1				
AT		4-6	0-2	0	6-12
WT		4-6	0-2	0	6-12
Total Items		**20**	**4**	**1**	
Total Points		**20**	**16**	**12**	**48**

Table 2.3.4
Actual Test Map for 2006 Grade 6 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC (Number of Items)	OE (Number of Items)	WT (Number of Items)	Total Points
Persuasive Prompt		0	0	1	12
Narrative	1				
AT		5	1	0	9
WT		5	1	0	9
Everyday Text	1				
AT		4	2	0	12
WT		6	0	0	6
Total Items		**20**	**4**	**1**	
Total Points		**20**	**16**	**12**	**48**

Table 2.3.5
Test Construction Map for Grade 7 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC (Number of Items)	OE (Number of Items)	WT (Number of Items)	Total Points
Persuasive Prompt		0	0	1	12
Narrative	1				
AT		4-6	0-2	0	6-12
WT		4-6	0-2	0	6-12
Everyday Text	1				
AT		4-6	0-2	0	6-12
WT		4-6	0-2	0	6-12
Total Items		**20**	**4**	**1**	
Total Points		**20**	**16**	**12**	**48**

Table 2.3.6
Actual Test Map for 2006 Grade 7 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC (Number of Items)	OE (Number of Items)	WT (Number of Items)	Total Points
Persuasive Prompt		0	0	1	12
Narrative	1				
AT		5	1	0	9
WT		5	1	0	9
Everyday Text	1				
AT		4	2	0	12
WT		6	0	0	6
Total Items		**20**	**4**	**1**	
Total Points		**20**	**16**	**12**	**48**

Table 2.3.7
Test Construction Map for Grade 5 Mathematics NJ ASK

Standard	Specified MC	Actual (1 pt.)	Specified OE	Actual (3 pts.)	Total Items	Total Points
I	6–10		0–1			
II	6–10		0–1			
III	6–10		0–1			
IV	6–10		0–1			
Total Items	**30**		**3**		**33**	
Total Points	**30**		**9**			**39**

Table 2.3.8
Actual Test Map for 2006 Grade 5 Mathematics NJ ASK

Standard	Specified MC	Actual (1 pt.)	Specified OE	Actual (3 pts.)	Total Items	Total Points
I	6–10	7	0–1	1	8	10
II	6–10	9	0–1	0	9	9
III	6–10	7	0–1	1	8	10
IV	6–10	7	0–1	1	8	10
Total Items	**30**	**30**	**3**	**3**	**33**	
Total Points	**30**	**30**	**9**	**9**		**39**

Table 2.3.9
Test Construction Map for Grade 6 Mathematics NJ ASK

Standard	Specified MC	Actual (1 pt.)	Specified OE	Actual (3 pts.)	Total Items	Total Points
I	6–10		0–1			
II	6–10		0–1			
III	6–10		0–1			
IV	6–10		0–1			
Total Items	**30**		**3**		**33**	
Total Points	**30**		**9**			**39**

Table 2.3.10
Actual Test Map for 2006 Grade 6 Mathematics NJ ASK

Standard	Specified MC	Actual (1 pt.)	Specified OE	Actual (3 pts.)	Total Items	Total Points
I	6–10	9	0–1	0	9	9
II	6–10	7	0–1	1	8	10
III	6–10	7	0–1	1	8	10
IV	6–10	7	0–1	1	8	10
Total Items	**30**	**30**	**3**	**3**	**33**	
Total Points	**30**	**30**	**9**	**9**		**39**

Table 2.3.11
Test Construction Map for NJ ASK Grade 7 Mathematics

Standard	Specified MC	Actual (1 pt.)	Specified OE	Actual (3 pts.)	Total Items	Total Points
I	6–10		0–1			
II	6–10		0–1			
III	6–10		0–1			
IV	6–10		0–1			
Total Items	**30**		**3**		**33**	
Total Points	**30**		**9**			**39**

Table 2.3.12
Actual Test Map for 2006 NJ ASK Grade 7 Mathematics

Standard	Specified MC	Actual (1 pt)	Specified OE	Actual (3 pts.)	Total Items	Total Points
I	6-10	7	0-1	1	8	10
II	6-10	9	0-1	0	9	9
III	6-10	7	0-1	1	8	10
IV	6-10	7	0-1	1	8	10
Total Items	**30**	**30**	**3**	**3**	**33**	
Total Points	**30**	**30**	**9**	**9**		**39**

2.4 Development of Test Items

The 2006 Grade 5, 6, and 7 NJ ASK consists of operational or base test items used to determine students' scores. These items originated from two sources:

1. Items field-tested on a former New Jersey Proficiency Assessment of State Standards (NJPASS) test but never used operationally. All Language Arts Literacy tests were constructed from this pool of items. Approximately one-half of the items used to construct the Mathematics tests were selected from this pool of items. All items were aligned to the current New Jersey Core Curriculum Content Standards.
2. Items field-tested on a former NJPASS test and used operationally. Approximately one-half of the items used to construct the Mathematics tests were selected from this pool of items. (For more detail, see Tables 2.9.1 through 2.9.3.) All items were aligned to the current New Jersey Core Curriculum Content Standards.

All items used on the 2006 NJ ASK went through the following steps of the item development process prior to 2000.

1. Riverside: Created test and item specifications based on New Jersey Core Curriculum Content Standards
2. Riverside: Selected and trained item writers
3. Item Writers: Wrote test items
4. Riverside: Conducted initial item review
5. Riverside: Conducted item review by experienced senior staff
6. Riverside: Conducted content and bias review with committees comprised of New Jersey educators
7. Field-tested items with New Jersey students (1998)
8. Riverside: Conducted Statistical Item Review

Some of these items were included on the NJPASS operational test. The remainder of the items were not used again in New Jersey until 2006.

In December 2005 and January 2006, the following additional development processes were undertaken.

9. Riverside: Aligned items to the New Jersey Core Curriculum Content Standards
10. NJDOE: Approved alignment of items, including the balance of standards reflected in the test blueprint; also improved item quality.
11. Removed all items that did not have NJDOE approval for adherence to the New Jersey Core Curriculum Content Standards.

2.5 Item Review Process

Riverside brought both field-tested non-used items and NJPASS operational items to New Jersey. Riverside reviewed the items with the NJDOE for strict adherence to the New Jersey Core Content Curriculum Standards. Any item that did not strictly adhere to an alignment was removed and replaced. Items were also reviewed for quality construct standards and data. Riverside and the NJDOE attempted to ensure that no changes were made to the items so that the data would not be affected. If the item did not meet the construct standard or was determined to have poor data, it was removed and replaced. No item used operationally was revised in such a way that it would have changed the alignment.

2.6 Item Use and Alignment

Tables 2.6.1 through 2.6.3 provide summary information on the alignment and usage history of items used on the 2006 NJ ASK operational tests.

Table 2.6.1
Grade 5 - Summary Information on Alignment and Usage History

	Number of Items		Max Score	Prev. OP Usage	
	MC	OE		Yes	No
LAL	**20**	**5**	**41**	**0**	**25**
Writing		1	5	0	1
Reading	20	4	36	0	24
Working with Text	12	1	16	0	13
Analyzing Text	8	3	20	0	11
Math	**30**	**3**	**39**	**16**	**17**
Number and Numerical Operation	7	1	10	4	4
Geometry and Measurement	9	0	9	3	6
Patterns and Algebra	7	1	10	4	4
Data Analysis, Probability, and Discrete Mathematics	7	1	10	5	3
Problem Solving	28	3	37	15	16

Table 2.6.2
Grade 6 - Summary Information on Alignment and Usage History

	Number of Items		Max Score	Prev. OP Usage	
	MC	OE		Yes	No
LAL	**20**	**5**	**48**	**0**	**25**
Writing		1	12	0	1
Reading	20	4	36	0	24
Working with Text	11	1	15	0	12
Analyzing Text	9	3	21	0	12
Math	**30**	**3**	**39**	**22**	**11**
Number and Numerical Operation	9	0	9	5	4
Geometry and Measurement	7	1	10	7	1
Patterns and Algebra	7	1	10	6	2
Data Analysis, Probability, and Discrete Mathematics	7	1	10	4	4
Problem Solving	28	3	37	21	10

Table 2.6.3
Grade 7 - Summary Information on Alignment and Usage History

	Number of Items		Max Score	Prev. OP Usage	
	MC	OE		Yes	No
LAL	**20**	**5**	**48**	**0**	**25**
Writing		1	12	0	1
Reading	20	4	36	0	24
Working with Text	12	0	12	0	12
Analyzing Text	8	4	24	0	12
Math	**30**	**3**	**39**	**14**	**19**
Number and Numerical Operation	7	1	10	3	5
Geometry and Measurement	9	0	9	3	6
Patterns and Algebra	7	1	10	5	3
Data Analysis, Probability, and Discrete Mathematics	7	1	10	3	5
Problem Solving	28	3	37	13	18

2.7 Test Forms Assembly

There are four steps associated with assembling test forms for NJ ASK:

1. Determine form design
2. Select items that meet content specifications
3. Evaluate statistical specifications and select items to meet these specifications
4. Review and approve test forms

Determine forms design – Each form consists of a set of operational items.

Select items that meet content specifications – Each content area measures subsets of items called clusters. In Language Arts Literacy the clusters include: Writing (Writing about Pictures [grade 5] and Writing about Persuasive Prompts [grades 6 and 7]) and Reading (Working with Text and Analyzing Text). In Mathematics the clusters include Number Sense and Numerical Operations; Geometry and Measurement; Patterns and Algebra; and Data Analysis, Probability, and Discrete Mathematics. There is also a process cluster called Problem Solving.

Evaluate statistical specifications and select items to meet these specifications – As forms are created it is necessary to determine if the statistical specifications have been met. Spreadsheets (form matrices) are used to provide information on the statistical properties of newly created forms. These matrices contain the following statistics (among others): Average p-value, biserial correlation and average IRT difficulty. Riverside's research group and the NJDOE will conduct an internal review of the data from year to year to ensure that test difficulty is equivalent. Riverside is prepared to utilize a set of anchor items that will remain constant to help with the calibration process.

Review and approve test forms – Once the content and statistical specifications have been met for each grade and subject, the forms are approved by the NJ DOE. The forms are then released for production and editorial reviews.

Checklists and quality control procedures accompany each stage of form development. Some of these procedures are listed on the following page:

2.8 Quality Control for Test Construction

Following is a list of quality control procedures used during the assembly of NJ ASK forms:

- Construct forms based on all content requirements noted in the test blueprint.
- Verify correct number of items per standard or reporting category based on test blueprint.
- Review selected items to ensure a wide sampling of the knowledge and skills being measured.
- Ensure that all selected items have been through the appropriate review procedures and are approved for use by the NJDOE.
- Check for a variety of item topics, equal distribution of male/female, ethnicities, etc.
- Verify appropriate portions of items with and without artwork.
- Check for cueing across all items on each form.
- Verify match of unique item identification numbers (UIN) to test matrix.
- Verify equal or nearly equal distribution of answer choices for MC items.
- Ensure that the test meets the statistical specifications, i.e., that as many items as possible have p-values between .35 and .9; as many items as possible have point biserials above .20; and the average Rasch is between 495 and 500 (see Appendix 2-1).
- Consider any statistical flags or problems.
- Check statistics to ensure that the collection of items yields an overall difficulty that falls within the specified range.
- Verify that items have not been released to the public.
- Verify correct answer key for each item.
- Perform content review of form (senior staff).
- Perform statistical review of form (psychometrician/statistician).
- Send form to NJDOE for review and approval.

2.9 Number of Catalog NJPASS Items Used on the NJ ASK

Following the administration of the NJ ASK 5-7 assessments and prior to the completion of scoring, it was discovered that a number of operational test items had previously appeared in published, Riverside Publishing Company materials associated with its NJPASS product. With guidance from the state's Technical Advisory Committee (TAC), the NJDOE and Riverside conducted an analysis of this inadvertent item exposure with regard to possible implications for the validity of the NJ ASK 5-7 scores, and concluded that there may be a small effect due to item exposure. A summary of that analysis follows. [4]

[4] For the full report, see *Revised Analysis of NJPASS Item Exposure for the NJ ASK Mathematics Assessment*.

The following bullets outline the chronology of NJPASS item exposure and analyses:

- Beginning in 2001, Riverside developed and sold NJPASS to interested schools and districts in New Jersey.
- In the summer of 2005, the state issued an RFP that was later cancelled.
- In November 2005, Riverside provided an unsolicited proposal to assist the state in developing an April 2006 test. Because of the compressed timeline for development, Riverside proposed using items from the NJPASS item bank.
- In December 2005, Riverside embargoed sales of the NJPASS catalog product.
- During test construction, Riverside used items that were included on the operational NJPASS catalog assessment in order to obtain sufficient content coverage required for the April 2006 mathematics forms (no NJPASS items were needed for the Language Arts Literacy forms).
- In April 2006, a school district in New Jersey reported that students recognized some of the items on the 2006 math assessment as they were taking the test. It was discovered that NJPASS catalog tests had been used as a resource for training conducted in schools around the state. The organization providing the training had not purchased these materials from Riverside, but had obtained them and copied them without permission for their workshops.
- In April after the tests had been administered across the state, the school district conveyed the above information to the NJDOE and faxed the training materials to the state department. Riverside began to research the organization to find out what other districts could have been exposed to the materials, but the level of exposure was unclear. No other district in the state reported this issue to the NJDOE or Riverside Publishing.
- Also in April, Riverside drafted a letter to the NJDOE that included a summary of the item overlap with published NJPASS materials and a preliminary opinion of how the item exposure could possibly affect student performance on the assessment. This document was shared with New Jersey's TAC.
- In June 2006, the TAC responded to the initial letter from Riverside and asked that a series of analyses be conducted by Riverside in order to measure the impact the item exposure may have had on student results. Riverside responded with proposed analyses and a timeline. The original item exposure report was presented to the TAC in July of 2006.
- At the July 2006 TAC meeting, the TAC had further recommendations and asked for additional analyses to be conducted. Riverside then went back and completed the additional analyses and presented an updated report at the November 2006 TAC meeting. This summary of the report represents the November 2006 report with the addition of several more analyses recommended by the TAC in November.

The number of previously exposed "on-grade-level" mathematics items varied by test. In the Grade 5 NJ ASK form, there were 10 items from the NJPASS, which constituted about 30% of the items. In the Grade 6 NJ ASK form, there were 8 items from NJPASS, which constituted about 24% of the items. Finally, in the Grade 7 NJ ASK form, there were 11 items from NJPASS, which constituted about 33% of the items. The number of previously exposed "off-grade-level" mathematics items also varied by test.

Tables 2.9.1 through 2.9.3 below provide more detail regarding the number of NJPASS items included on the NJ ASK forms. These tables contain information on the number and percentage of mathematics items that were used on the 2006 NJ ASK that were also on the published forms of NJPASS (the percentages may not add up exactly because of rounding). The NJ ASK mathematics tests at grades 5–7 all have 30 multiple-choice and 3 constructed-response items, for a total of 33 items and 39 points. An asterisk (*) by the NJPASS form in the table indicates the published form that corresponds to the same grade level as the NJ ASK assessment.

Table 2.9.1
Grade 5 - The Number of NJPASS Items Included on the 2006 NJ ASK

2006 NJ ASK Grade 5			
Number of Items from NJPASS 56*	Number of Items from NJPASS 67	Number of Items from NJPASS 78	Number of Items from NJ PASS
10	3	3	16
Percentage of Items from NJPASS 56*	Percentage of Items from NJPASS 67	Percentage of Items from NJPASS 78	Percentage of Items from NJ PASS
30%	9%	9%	48%
Percentage of Total Points from NJPASS 56*	Percentage of Total Points from NJPASS 67	Percentage of Total Points from NJPASS 78	Percentage of Total Points from NJPASS
26%	18%	8%	51%

Table 2.9.2
Grade 6 - The Number of NJPASS Items Included on the 2006 NJ ASK

2006 NJ ASK Grade 6			
Number of Items from NJPASS 56	Number of Items from NJPASS 67*	Number of Items from NJPASS 78	Number of Items from NJ PASS
8	8	6	22
Percentage of Items from NJPASS 56	Percentage of Items from NJPASS 67*	Percentage of Items from NJPASS 78	Percentage of Items from NJ PASS
24%	24%	18%	67%
Percentage of Total Points from NJPASS 56	Percentage of Total Points from NJPASS 67*	Percentage of Total Points from NJPASS 78	Percentage of Total Points from NJPASS
21%	21%	15%	56%

Table 2.9.3
Grade 7 - The Number of NJPASS Items Included on the 2006 NJ ASK

2006 NJ ASK Grade 7			
Number of Items from NJPASS 56	Number of Items from NJPASS 67	Number of Items from NJPASS 78*	Number of Items from NJ PASS
1	2	11	14
Percentage of Items from NJPASS 56	Percentage of Items from NJPASS 67	Percentage of Items from NJPASS 78*	Percentage of Items from NJ PASS
3%	6%	33%	42%
Percentage of Total Points from NJPASS 56	Percentage of Total Points from NJPASS 67	Percentage of Total Points from NJPASS 78*	Percentage of Total Points from NJPASS
3%	5%	28%	36%

Approximately 10% of New Jersey school districts either purchased the NJPASS or had a school within the district directly purchasing the NJPASS. More specifically, about 30 of 590 districts at each grade level (i.e., grades 5, 6, and 7), plus approximately 40 individual schools per grade level from these and various other districts, purchased NJPASS. When all districts having at least one exposed school were tabulated, the number of districts exposed to NJPASS items at grades 5, 6, and 7 were 56, 56, and 54, respectively. When the number of schools with possible exposure was tabulated (both individual school purchases plus schools within districts that purchased NJPASS), the counts at grades 5, 6, and 7 were 144, 93, and 82, respectively.

Item Exposure Analyses and Results

The analyses presented below were based on the calculation of four "new" scale scores. However, before calculating these new scale scores, the 2006 NJ ASK operational test was recalibrated using only the school districts that were *not* exposed to NJPASS items[5]. This recalibration was recommended by New Jersey's Technical Advisory Committee (TAC), and as noted above, approximately 90% of school districts were not exposed to NJPASS items.

After recalibration, the four new scale scores were calculated for each grade level of NJ ASK. Scale scores were generated from each of the following NJ ASK item subsets:

a) NJ ASK items that were **not included** in NJPASS
b) NJ ASK items that were **on-grade level** NJPASS items
c) NJ ASK items that were **off-grade level** NJPASS items, and
d) NJ ASK items that were either on-grade or off-grade level NJPASS items (**combined**).

One key recommendation proposed by New Jersey's TAC was to focus the analysis of possible exposure at the school level rather than at the district level. This refocusing of the analysis at the school level was undertaken to address possible confounding/masking of

[5] A customer list of districts and individual schools that had purchased NJPASS in either 2004 or 2005 was used to identify exposed school districts. The only exception was a single district that did not directly purchase NJPASS; however, they were exposed to NJPASS materials.

effects in the initial analyses that originally targeted districts that had access to NJPASS materials. The analyses described below modified decision rules to better isolate schools likely to have seen NJPASS materials.

For each grade, schools were categorized into one of three Groups:

1) relatively clear indication of exposure (direct school purchase)
2) possible exposure (district purchase with supplies sufficient for all schools)[6]
3) no indication of exposure

Comparisons of the above Exposure Groups provided an indication of possible exposure effects. More specifically, comparisons of mean scale scores between Groups 1 and 3 and between Groups 2 and 3 were examined. In addition to comparing mean scale scores across the three groups, an analysis of data from matched samples was performed. For each grade level, obtaining the matched samples included the following steps:

- determining the variables to use in matching, including the granularity of matching values,
- determining the matching process, including criteria for handling multiple matches for a single "target" school and aggregating the matched data, and
- determining the overall impact of item exposure at the scale score level and the proficiency classification level.

An analysis using matched samples comparing NJPASS schools and non-NJPASS schools was conducted using the following variables:

a) New Jersey's District Factor Groups (DFG). New Jersey has a long history of using DFGs in the analysis of assessment results. The statistical method for developing DFGs includes U.S. Census data and has been improved over the years to provide a better measure of social economic status (SES). The eight DFGs used for matching schools included A, B, CD, DE, FG, GH, I, and J.

b) Language Arts Literacy (LAL) performance. New Jersey's TAC specifically recommended that LAL performance be considered when studying differences between exposed and unexposed schools. Incorporating LAL performance into the matching methodology allowed for the implementation of this recommendation. The following score ranges were used for matching *average* school performance:

   - LAL Category 1 = Mean LAL Scale Score ≤ 200
   - LAL Category 2 = Mean LAL Scale Score > 200 and ≤ 210
   - LAL Category 3 = Mean LAL Scale Score > 210 and ≤ 220
   - LAL Category 4 = Mean LAL Scale Score > 220 and ≤ 230
   - LAL Category 5 = Mean LAL Scale Score > 230

[6] This group also included the one district that did not directly purchase NJPASS; however, they were exposed to NJPASS materials.

c) School size. Because an indication of school size is no longer included in DFG determinations, and because of the wide range of school sizes in New Jersey, the following "school" size categories were used in matching. Note that the size variable was based on the total number of students within a specific grade (i.e., either grade 5, 6 or 7).

- Size Category 1 = Mean frequency count for a grade ≤ 100
- Size Category 2 = Mean frequency count for a grade > 100 and ≤ 200
- Size Category 3 = Mean frequency count for a grade > 200 and ≤ 300
- Size Category 4 = Mean frequency count for a grade > 300 and ≤ 400
- Size Category 5 = Mean frequency count for a grade > 400

The above variables were created for each school in each Exposure Group by grade level. Exposure Group 1 (Yes – School Exposed) was then matched with similar schools in Exposure Group 3 (No – School Not Exposed). The same process was used for matching Exposure Group 2 schools (Yes – School Exposed in District) with similar schools in Exposure Group 3 (No – School Not Exposed). Table 2.9.4 presents the number of matching schools in each Exposure Group by grade level.

The large number of Exposure Group 3 schools as compared to Groups 1 and 2 allowed for multiple samples of matched Group 3 schools to be selected. Multiple samples were drawn with replacement by using a uniform random number generator. All together, 10 different samples (i.e., replications) of schools that matched the Group 1 and Group 2 schools were selected. Results for the matched data were aggregated across the 10 replications.

Table 2.9.4
2006 NJ ASK Item Exposure Analysis
Number of Matching Schools in Each Exposure Group by Grade Level

Grade	Exposure Group	Number of Schools	Number Matching	Percent Matching
5	1. Yes – School Exposed	38	36	94.7%
	2. Yes – School Exposed in District	106	103	97.2%
	3. No – School Not Exposed	1054	N/A	N/A
6	1. Yes – School Exposed	40	37	92.5%
	2. Yes – School Exposed in District	53	53	100.0%
	3. No – School Not Exposed	728	N/A	N/A
7	1. Yes – School Exposed	30	28	93.3%
	2. Yes – School Exposed in District	52	49	94.2%
	3. No – School Not Exposed	592	N/A	N/A

Based upon the results in Tables 2.9.5 and 2.9.6, using the Adjusted Difference[7], the impact on NJ ASK mathematics scale scores appears to be the *addition* of approximately 1.0 scale score point at grade 5, 2.0 scale score points at grade 6, and 2.5 scale score points at grade 7. It is important to note, however, that these additional points apply to scale scores based on only one-half of the total item set (see Tables 2.9.1 through 2.9.3). Because there were no exposure issues for the remaining one-half of the total item set, the scale score impact just noted (i.e., 1.0, 2.0 and 2.5 for grades 5, 6, and 7, respectively) may be extreme and could arguably be cut in half. While there appears to be a complex interaction (or intermingling) of school effect and item effect, some "diluting" of item effect above and beyond the Adjusted Difference would probably occur when exposed items are combined with unexposed items to create operational scores. For the purposes of discussing impact, however, the above values will be used.

The impact on student performance classifications can be ascertained by viewing Tables 7-1 through 7-3 in Appendix 7-1. Tables 7-1 through 7-3 contain the raw score to scale score conversions for the 2006 NJ ASK operational mathematics assessments for grades 5, 6, and 7, respectively. Raw score cut points and their accompanying scale score are highlighted for both the Proficiency (scale score of 200) and Advanced Proficiency (scale score 250) performance classifications. As can be seen from these tables, the scale score difference from the raw score cut point down to the next raw score (i.e., minus one raw score point) is never less than 3 scale score points. In other words, if 2.5 scale score points (recall that this is probably an extreme estimate of the impact on scale scores) were added to every student's obtained mathematics scale score, this amount would *not* be large enough to push students who were just below the cut point into the next performance category. For example, for the grade 5 operational scale (Table 7-1), the student would need 4 scale score points to move into the Proficiency classification. This observation suggests that while there may be a small effect due to item exposure, it is not large enough to have much of an impact on student proficiency classifications.

To further study the impact on student proficiency classifications, an analysis of student pass rates was conducted. More specifically, the percentage of students scoring proficient and above was compared for both mathematics and LAL for matched exposed and not exposed schools. Tables 2.9.7 and 2.9.8 present the results comparing matched exposed and not exposed schools for percent proficient (percent passing) based on the total scale score for either mathematics or LAL. In general, differences between the passing rates for exposed and not exposed schools were larger for mathematics than for LAL. For mathematics, half of the observed differences were less than one percentage point while the remaining differences were somewhat larger. This finding is consistent with the above observation that there may be an effect due to item exposure, but that effect is small and inconsistent across grades.

[7] The Adjusted Difference takes the difference seen with exposed items and subtracts out the difference seen with unexposed items. The adjusted difference was calculated to take into account the difference in "ability" or "achievement" observed between the exposed schools and non-exposed schools, even after matching. Calculating the Adjusted Difference was recommended by New Jersey's TAC and possibly represents the best estimate of the impact of exposure on scale scores derived from only the exposed items.

Table 2.9.5
2006 NJ ASK Item Exposure Analysis
Results Comparing Matched Exposed and Not Exposed Schools
For Scale Scores Based on Non Exposed Items or Exposed Items
Comparison Between 1. Yes – School Exposed and 3. No – School Not Exposed

	Items Exposed	School Exposed	School N-Count	Mean	Difference	Adjusted Difference**
Grade 5	No (17 items)	No (Grp. 3)	36*	224.2	-2.7	-1.1
		Yes (Grp. 1)	36	226.9		
	Combined (16 items)	No (Grp. 3)	36*	225.0	-3.8	
		Yes (Grp. 1)	36	228.8		
Grade 6	No (11 items)	No (Grp. 3)	37*	215.1	-1.3	-2.2
		Yes (Grp. 1)	37	216.4		
	Combined (22 items)	No (Grp. 3)	37*	216.1	-3.5	
		Yes (Grp. 1)	37	219.6		
Grade 7	No (19 items)	No (Grp. 3)	28*	213.2	-0.4	-2.6
		Yes (Grp. 1)	28	213.6		
	Combined (14 items)	No (Grp. 3)	28*	214.1	-3.0	
		Yes (Grp. 1)	28	217.1		

* Average over 10 replications of School N-Count.
** Difference between groups on Combined items (e.g., -3.8) and groups on Non-exposed items (e.g., -2.7).

Table 2.9.6
2006 NJ ASK Item Exposure Analysis
Results Comparing Matched Exposed and Not Exposed Schools
For Scale Scores Based on Non Exposed Items or Exposed Items
Comparison Between 2. Yes – School Exposed in District and
3. No – School Not Exposed

	Items Exposed	School Exposed	School N-Count	Mean	Difference	Adjusted Difference**
Grade 5	No (17 items)	No (Grp. 3)	103*	229.7	-1.7	-0.3
		Yes (Grp. 2)	103	231.4		
	Combined (16 items)	No (Grp. 3)	103*	230.8	-2.0	
		Yes (Grp. 2)	103	232.8		
Grade 6	No (11 items)	No (Grp. 3)	53*	212.4	-3.3	-0.3
		Yes (Grp. 2)	53	215.7		
	Combined (22 items)	No (Grp. 3)	53*	214.3	-3.6	
		Yes (Grp. 2)	53	217.9		
Grade 7	No (19 items)	No (Grp. 3)	49*	210.7	0.5	-1.8
		Yes (Grp. 2)	49	210.2		
	Combined (14 items)	No (Grp. 3)	49*	211.1	-1.3	
		Yes (Grp. 2)	49	212.4		

* Average over 10 replications of School N-Count.
** Difference between groups on Combined items (e.g., -2.0) and groups on Non-exposed items (e.g., -1.7).

Table 2.9.7
2006 NJ ASK Item Exposure Analysis
Results Comparing Matched Exposed and Not Exposed Schools
For Percent Proficient (Percent Passing) Based on Total Scale Score
Comparison Between 1. Yes – School Exposed and 3. No – School Not Exposed

	School Exposed	School N-Count	Percent Proficient Math	Difference Math	Percent Proficient LAL	Difference LAL
Grade 5	No (Grp. 3)	36*	79.8	-2.5%	84.3	-0.4%
	Yes (Grp. 1)	36	82.3		84.7	
Grade 6	No (Grp. 3)	37*	70.6	-0.8%	74.4	0.3%
	Yes (Grp. 1)	37	71.4		74.1	
Grade 7	No (Grp. 3)	28*	67.9	-0.2%	83.5	-0.1%
	Yes (Grp. 1)	28	68.1		83.6	

* Average over 10 replications of School N-Count.

Table 2.9.8
2006 NJ ASK Item Exposure Analysis
Results Comparing Matched Exposed and Not Exposed Schools
For Percent Proficient (Percent Passing) Based on Total Scale Score
Comparison Between 2. Yes – School Exposed in District and
3. No – School Not Exposed

	School Exposed	School N-Count	Percent Proficient Math	Difference Math	Percent Proficient LAL	Difference LAL
Grade 5	No (Grp. 3)	36*	84.7	-1.1%	88.3	0.5%
	Yes (Grp. 2)	36	85.8		87.8	
Grade 6	No (Grp. 3)	37*	67.6	-5.1%	73.8	1.0%
	Yes (Grp. 2)	37	72.7		72.8	
Grade 7	No (Grp. 3)	28*	62.8	0.5%	79.6	1.5%
	Yes (Grp. 2)	28	62.3		78.1	

* Average over 10 replications of School N-Count.

2.10 Summary

The NJ ASK for grades 5–7 provides an indication of student progress toward achieving the knowledge and skills identified in the Core Curriculum Content Standards (CCCS) and the tests fulfill the requirements under NCLB. Just as the CCCS guided the item development and selection process, the consideration of content played an equally important role in forms development. Forms development required a balance of both content coverage and item difficulty. As items were selected for inclusion on particular forms, every effort was made to balance the content coverage to ensure they aligned to the CCCS being assessed while simultaneously considering the difficulty of the forms.

PART 3: STANDARD SETTING

3.1 Introduction

The 2006 NJ ASK assessed two subject areas, Language Arts Literacy (LAL) and mathematics, in grades 5, 6, and 7. Each subject area included both multiple-choice items and open-ended items. After the April 2006 administration, a standard setting workshop was held to determine the cut scores for LAL and mathematics that would distinguish performance among three levels: Partially Proficient, Proficient, and Advanced Proficient.

The NJ ASK 2006 grades 5–7 standard setting had two phases. Phase 1 was a two-day session and was held on May 11 and 12, 2006. It involved 118 educators from across the state of New Jersey (see Appendix 3-1 for the demographic background of panelists) and used a research-based standard setting method to recommend cut scores. The Phase 2 meeting was held on May 15, 2006. It included one person from each of the Phase 1 panels as well as five additional policymakers from the state. A state board meeting was held on May 17, 2006, in Trenton, and the recommended cut scores from Phase 2 were approved.

The full Standard Setting report, available from the New Jersey Department of Education (NJDOE), provides details about the standard setting procedures, demographic information of the panelists, panelists' ratings from one round to the next, and their responses on the evaluation forms. The final cut scores approved by the State Board of Education are also presented. The sections below summarize the most important steps of the standard setting process. For more detail, the full Standard Setting Report should be referenced.

3.2 Development of Performance Level Descriptors

On April 19, 2006, a group of 57 educators participated in a one-day workshop to develop the Performance Level Descriptors (PLDs) for Proficient and Advanced Proficient in LAL and mathematics. These descriptors were used throughout the standard-setting process and can found in Appendix 3-2. The NJ ASK 2006 grades 5–7 PLD meeting included the following steps:

1. Orientation - All panelists gathered as one large group to view a Power Point presentation prepared by the NJDOE. The presentation discussed the development of NJ ASK 2006 grades 5–7 and apprised the panelists as to what a PLD is and how it would be used by panelists in Phase 1 of standard setting to develop cut scores. For the day's activities each panelist was given a copy of New Jersey's Core Curriculum Content Standards (CCCS).
2. Panelists then broke up into six preplanned groups - one per content area and grade level. Each panelist then took the appropriate NJ ASK 2006 grades 5–7 operational test.

3. Panelists were provided time to discuss the test, including its content, challenges, and difficulty.
4. Panelists divided into smaller groups and made lists of cognitive characteristics required by the NJ CCCS that would define Proficient and Advanced Proficient for their assigned grade and content area. This process was facilitated by NJDOE staff, and the documents and the prompts were supplied by the NJDOE.
5. All panelists returned to their grade and content group and the cognitive characteristics developed in Step 4 were merged into a narrative PLD. This process was facilitated by NJDOE staff, and the documents and the prompts were supplied by the NJDOE.
6. PLDs from all three grade's PLDs were read to all of the mathematics and language arts literacy panelists, and time was allowed for edits and suggestions from content appropriate panelists from other grades and NJDOE facilitators. PLDs were checked for vertical scaling during this time.
7. Panelists returned to grade assignments to consider and incorporate edits and suggestions to their PLDs.
8. Final PLDs were then presented to all appropriate content panelists.
9. Panelists completed a questionnaire developed by NJDOE.
10. PLDs were presented to all members of the Office of Evaluation and Assessment for edits and suggestions. No edits or suggestions were made and the PLDs that were developed on April 19, 2006, were used on May 11 and 12 for Standard Setting Phase 1.

3.3 Standard Setting Process

The NJ ASK 2006 grades 5–7 standard setting included student data based on a sample of early returned data that consisted of about 14% of the student population. The scores for 42,569 students from grades 5, 6, and 7 were used in setting the LAL standards, and a total of 42,592 students were used in setting the mathematics standards.

Since the LAL and mathematics tests included both multiple-choice and open-ended items, a modified-Angoff method (Hambleton & Plake, 1995)[8] was implemented. The Angoff method is the most thoroughly researched method used in setting standards. According to Loomis & Bourke (2001)[9], "No other method could be identified that appeared to be as easy to use, as technically sound, and as well researched as the Angoff method."

[8] Hambleton, R. K. & Plake, B. S (1995) Using an Extended Angoff Procedure to Set Standards on Complex Performance Assessments. *Applied Measurement in Education,* **8**, 1, 41-55.
[9] Loomis, S. C. & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. J. Cizek (Ed.), Standard performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Lawrence Erlbaum Associates.

Phase 1

The NJ ASK 2006 grades 5–7 standard-setting procedures contained two phases. Phase 1 included the following 10 steps:

1. Orientation to the NJ ASK 2006 program and the purpose and use of standards.
2. Administration of the test to each judge.
3. Discussion and operational definition of the *Proficient* and *Advanced Proficient* Performance Level Descriptors.
4. Initial, independent ratings by judges, each estimating the number of students out of 100 who would answer each multiple-choice test question correctly. Judges were provided forms (e.g., Standard-Setting Item Rating Form) for accomplishing this task.
5. Data tabulation for each judge, and across averages across all judges, to derive the mean cut score for the panel.
6. Discussion of initial estimates derived from the Stage 1 ratings and review of item-level information (p-values) based on actual student data from the early return sample.
7. Stage 2 ratings occurred after results were reviewed and the new information was discussed. Each judge had an opportunity to change or reaffirm his/her decision regarding each item.
8. Discussion of the overall judges' Stage 2 cut scores for the two standards of *Proficient* and *Advanced Proficient*.
9. Final estimates were determined after presentation and discussion of the Stage 2 data. Based on these Stage 3 ratings, recommendations were made for *Proficient* and *Advanced Proficient* performance.
10. Evaluation of the standard-setting process by each judge.

Phase 2

The Phase 2 meeting took place on May 15, 2006, three days after the modified-Angoff event. One judge from each of the Phase 1 panels and five policymakers, including some superintendents and other stakeholders, participated in the meeting.

The purpose of the Phase 2 panel was to look at the big picture of the standards across grades and subjects. Their responsibility was to review the Phase 1 results from a policy point of view and make any changes that were necessary from this viewpoint. The Phase 2 panelists reviewed the results and considered consistency across all grades (grades 3 through 8) and assessed content areas in terms of both the percentage of points required to reach each performance level and the percentage of students reaching each performance level. They were provided information on how the initial cut scores were determined, the value of the cut scores, and the projected impact in the student population. Phase 2 panelists were also given graphs which showed the percentage of students in each classification group for Grades 3, 4, and 8, as well as the percentage of students in each classification group developed by the Grade 5–7 standard setting.

The primary task of Phase 2 panelists was to determine the reasonableness of the standards from a policy perspective and to recommend adopting either the Phase 1 cut scores or another set of cut scores within reasonable limits of the originals (e.g., up to +/- 4-point range).

Phase 2 of NJ ASK 2006 grades 5–7 standard setting included the following:

1. Choose the Panel. The panel was composed of six individuals consisting of a subset of the modified-Angoff panelists supplemented by five state and local policymakers, including district superintendents and curriculum supervisors. The modified-Angoff panelists were recommended from each of the six panels. An attempt was made to select modified-Angoff panelists who the NJDOE and RPC felt could best represent their respective panel in terms of having the ability to communicate the groups' thoughts and discussions. In addition, modified-Angoff panelists provided representation from both genders and various ethnic and DFG groups.
2. Summarize Phase 1 Procedure. It was important to ensure that the Phase 2 panelists had a complete understanding of the Phase 1 process. To this end, facilitators described the modified-Angoff procedures to the Phase 2 panelists so they could understand how the cut scores were determined. Participants from Phase 1 were asked to provide their own insights on important issues discussed in determining the Phase 1 cut scores.
3. Present Impact Data. The Phase 2 panelists were shown tabled results from the Phase 1 event, including impact data for the recommended cut scores. In addition, impact data for cut scores +/- 2 raw score points were also provided. Impact data were provided for all students as well as for important state and NCLB reporting groups, such as gender, ethnic group, educational programs, economical status, and migrant status. Panelists were also asked to consider the importance of consistency across grades and subjects.
4. Discuss Final Cut Scores. After panelists were trained on the methodology and saw the data described above, they discussed the results. Facilitators focused the panelists' discussion on whether the percentage of students reaching Proficient and Advanced Proficient were viable percentages in New Jersey.
5. Make Final Recommendations. After reviewing the Proficient and Advanced Proficient impact data and discussing the percentages both within and across grades levels, the Phase 2 panelists came to a decision on what cut scores to recommend. Cut scores recommended by Phase 2 panelists were submitted to the New Jersey State Board of Education.

3.4 Summary of Results

Overall, panelists' judgments about the cut scores converged from one round to the next, showing little variance by the end of the final round of Phase 1. This was reflected in the standard error of judgment (SEJ) that is related to the variance in panelists' judgments around the cut score. More specifically, the SEJ is the standard deviation calculated from the total

score based on the sum of the judges' p-value ratings across all test items. In general, the SEJs decreased across rounds, indicating converging opinions. See the RESULTS section of the full Standard Setting Report (page 27) for more details.

The NJ ASK 2006 grades 5, 6, and 7 cut scores and the corresponding conditional standard error of measurement (CSEM) approved by the New Jersey State Board of Education on May 17, 2006, are summarized in Table 3.4.1. The CSEM is calculated as:

$$CSEM = (SD_{obs} / SD_{adj'd}) * SE_{theta}$$

Where: SD_{obs} is the observed standard deviation for number correct raw score,
$SD_{adj'd}$ is an estimate of the "true" sample standard deviation, and
SE_{theta} is the standard error for theta.

Table 3.4.1
NJ ASK 2006 Grades 5–7 Cut Scores with Conditional Standard Error of Measurement by Content Area and Grade Level

		Language Arts Literacy		Mathematics	
		Proficient	Advanced Proficient	Proficient	Advanced Proficient
Grade 5	Cut score*	16	30	18	30
	CSEM	1.7	2.7	2.5	2.7
Grade 6	Cut score*	20	32	17	31
	CSEM	1.5	2.0	3.8	4.4
Grade 7	Cut score*	21	34	13	26
	CSEM	1.5	2.2	3.4	3.2

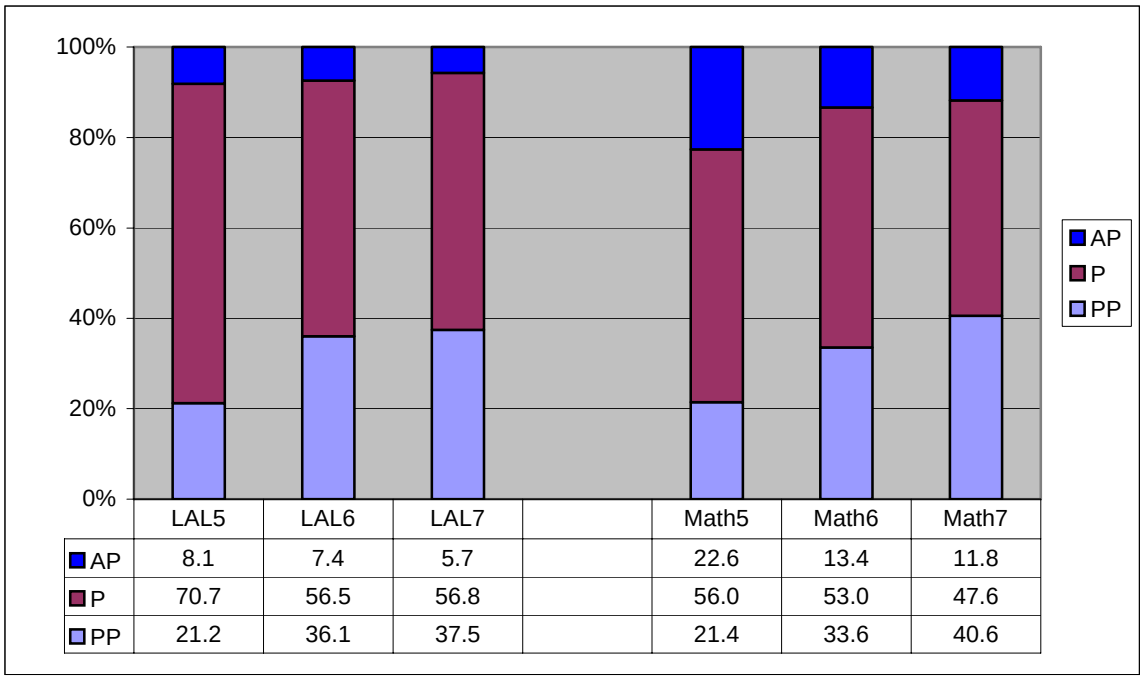
*Cut scores were approved by the New Jersey State Board of Education on May 17, 2006.

Distributions of students in each proficient category—Partially Proficient (PP), Proficient (P), and Advanced Proficient (AP)—are tabulated in Table 3.4.2 and presented graphically in Figure 3.4.1.

Table 3.4.2
Distribution of Students in Each Proficient Category and Test Characteristics by Content Area and Grade Level

	Grade	N-count*	% Partially Proficient	% Proficient	% Advanced Proficient	N-item	Mean Raw Score	Maximum Possible Score
LAL	5	14181	21.2	70.7	8.1	25	20.80	41
	6	14206	36.1	56.5	7.4	25	21.64	48
	7	14182	37.5	56.8	5.7	25	22.42	48
Math	5	14185	21.4	56.0	22.6	33	23.41	39
	6	14208	33.6	53.0	13.4	33	20.53	39
	7	14199	40.6	47.6	11.8	33	15.57	39

*Based on early returned sample data.



	LAL5	LAL6	LAL7		Math5	Math6	Math7
AP	8.1	7.4	5.7		22.6	13.4	11.8
P	70.7	56.5	56.8		56.0	53.0	47.6
PP	21.2	36.1	37.5		21.4	33.6	40.6

Figure 3.4.1: Distribution of Students in Each Proficient Category by Content Area and Grade Level

3.5 State Board of Education Review and Adoption

A New Jersey State Board of Education meeting was held May 17, 2006, two days after the Phase 2 meeting. An introduction was first presented to the Board members. The introduction included a definition of the concept of "Standard Setting," NJ ASK proficiency levels, the Angoff procedure, and validity considerations. Following that, the methods of rating and calculating the cut scores were demonstrated. Phase 1 and Phase 2 results were presented. In addition, distributions of students by subgroups, such as gender, ethnic group, and educational programs, were presented. Validity evidence of the standard setting procedures as well as the cut scores were also presented. All cut scores recommended by the Phase 2 panelists were approved by the State Board of Education on May 17, 2006.

PART 4: ITEM ANALYSIS

4.1 Introduction

Item analyses were conducted for the 2006 NJ ASK for grades 5, 6, and 7 in the content areas of Language Arts Literacy (LAL) and Mathematics. In this section, we present summary information by grade at both the content domain and content cluster level. The information includes mean item scores and discrimination indices. The data are based on 104,390, 105,265, and 108,372 students, for grades 5, 6, and 7, respectively.

For multiple-choice (MC) items, the mean score is simply the proportion of students who gave a correct response to the item (usually referred to as item difficulty or the p-value), and the discrimination index is the point biserial correlation between the item score and the total score based on the remaining items. For LAL, the test criterion score was the total score of all reading items (MC and OE) and the writing prompt. For mathematics, the test criterion score was the total score of all MC and OE items.

For open-ended (OE) items, the mean score is the mean of students' scores on a scale of 0 to 3 for mathematics and 0 to 4 for LAL. Writing is scored on a scale of 0 to 5 for grade 5 and 0 to 6 for grades 6 and 7. Note that for grades 6 and 7, the writing scores were doubled in data analyses and score reporting. The discrimination index is the correlation between the item score and the total score based on the remaining items.

4.2 Grade 5, 6 and 7 Forms

Tables 4.2.1, 4.2.2, and 4.2.3 summarize, by item response format, item difficulty and discrimination of the items that comprise each content domain and cluster for grades 5, 6, and 7 respectively. For MC items, both the mean and standard deviation are given. The mean value is the average of the p-values of the items in the cluster. For OE items, the mean value is the average item score for the items in the cluster. Item discrimination is the correlation between students' item score and the total score of the remaining items on the test. Both item difficulty and discrimination are expressed in terms of the raw score metric.

Tables 4.2.4, 4.2.5, and 4.2.6 summarize frequency distributions for MC item difficulty and discrimination indices of items comprising each content domain and cluster for grades 5, 6, and 7, respectively. The median item difficulty and discrimination is also displayed.

Table 4.2.7 summarizes distractor analyses of MC items by test form. The number in each cell indicates the number of items that at least one p-value or discrimination index (point-biserial) of the distractors was higher than the keyed option.

Table 4.2.1
Grade 5 - Item Difficulty and Discrimination Summary Statistics for Multiple-Choice and Open-Ended Items by Content Area and Cluster

Test Section/ Cluster	Multiple-choice				Open-ended			
	Item Difficulty			Item Discrimination	Item Difficulty			Item Discrimination
	Nitem	Mean	S.D.	Mean	Nitem	Mean	S.D.	Mean
LAL	**20**	**0.71**	**0.14**	**0.39**	**5**	**1.62**	**0.76**	**0.50**
Writing					1	2.93		0.50
Reading	20	0.71	0.14	0.39	4	1.30	0.25	0.50
Interpreting Text	12	0.64	0.12	0.38	1	1.55		0.54
Analyzing Text	8	0.82	0.07	0.40	3	1.21	0.23	0.48
Math	**30**	**0.67**	**0.19**	**0.34**	**3**	**1.47**	**0.27**	**0.52**
Number and Numerical Operation	7	0.75	0.15	0.36	1	1.19		0.54
Geometry and Measurement	9	0.67	0.20	0.27				
Patterns and Algebra	7	0.57	0.23	0.35	1	1.49		0.48
Data Analysis, Probability, and Discrete Mathematics	7	0.67	0.15	0.40	1	1.73		0.54
Problem Solving	28	0.66	0.19	0.33	3	1.47	0.27	0.52

Table 4.2.2
Grade 6 - Item Difficulty and Discrimination Summary Statistics for Multiple-Choice and Open-Ended Items by Content Area and Cluster

Test Section/ Cluster	Multiple-choice				Open-ended			
	Item Difficulty			Item Discrimination	Item Difficulty			Item Discrimination
	Nitem	Mean	S.D.	Mean	Nitem	Mean	S.D.	Mean
LAL	**20**	**0.68**	**0.14**	**0.32**	**5**	**1.97**	**1.84**	**0.47**
Writing					1	5.24		0.55
Reading	20	0.68	0.14	0.32	4	1.15	0.20	0.45
Interpreting Text	11	0.66	0.15	0.34	1	1.25		0.48
Analyzing Text	9	0.71	0.13	0.30	3	1.11	0.23	0.44
Math	**30**	**0.60**	**0.13**	**0.39**	**3**	**1.27**	**0.53**	**0.58**
Number and Numerical Operation	9	0.61	0.18	0.36				
Geometry and Measurement	7	0.68	0.10	0.41	1	1.19		0.61
Patterns and Algebra	7	0.55	0.08	0.39	1	0.79		0.56
Data Analysis, Probability, and Discrete Mathematics	7	0.55	0.11	0.41	1	1.83		0.58
Problem Solving	28	0.59	0.12	0.39	3	1.27	0.53	0.58

Table 4.2.3
Grade 7 - Item Difficulty and Discrimination Summary Statistics for Multiple-Choice and Open-Ended Items by Content Area and Cluster

Test Section/ Cluster	Multiple-choice				Open-ended			
	Item Difficulty			Item Discrimination	Item Difficulty			Item Discrimination
	Nitem	Mean	S.D.	Mean	Nitem	Mean	S.D.	Mean
LAL	**20**	**0.73**	**0.12**	**0.34**	**5**	**2.23**	**1.85**	**0.51**
Writing					1	5.54		0.54
Reading	20	0.73	0.12	0.34	4	1.40	0.07	0.51
Interpreting Text	12	0.71	0.11	0.33				
Analyzing Text	8	0.76	0.13	0.37	4	1.40	0.07	0.51
Math	**30**	**0.48**	**0.15**	**0.34**	**3**	**0.71**	**0.15**	**0.57**
Number and Numerical Operation	7	0.54	0.10	0.36	1	0.65		0.55
Geometry and Measurement	9	0.48	0.20	0.31				
Patterns and Algebra	7	0.37	0.14	0.35	1	0.88		0.64
Data Analysis, Probability, and Discrete Mathematics	7	0.52	0.11	0.34	1	0.61		0.51
Problem Solving	28	0.47	0.15	0.34	3	0.71	0.15	0.57

Table 4.2.4
Grade 5 - Frequency Distributions for Multiple-Choice p-values and Discrimination Indices by Content Area and Cluster

	Nitem	p-value						Discrimination					
		Median	p < 0.25	0.25 <= p < 0.50	0.50 <= p < 0.75	0.75 <= p < 0.90	p >= 0.90	Median	*pb < 0.20	0.20 <= pb < 0.30	0.30 <= pb < 0.40	0.40 <= pb < 0.50	pb >= 0.50
LAL	**20**	**0.74**	**0**	**2**	**8**	**9**	**1**	**0.40**	**0**	**2**	**7**	**11**	**0**
Interpreting Text	12	0.65	0	2	8	2	0	0.39	0	1	6	5	0
Analyzing Text	8	0.79	0	0	0	7	1	0.42	0	1	1	6	0
Math	**30**	**0.69**	**1**	**6**	**12**	**9**	**2**	**0.36**	**2**	**5**	**15**	**8**	**0**
Number and Numerical Operation	7	0.79	0	1	2	3	1	0.36	0	1	4	2	0
Geometry and Measurement	9	0.70	0	2	3	4	0	0.25	2	3	3	1	0
Patterns and Algebra	7	0.53	1	2	2	2	0	0.33	0	0	6	1	0
Data Analysis, Probability, and Discrete Mathematics	7	0.68	0	1	5	0	1	0.44	0	1	2	4	0
Problem Solving	28	0.69	1	6	11	8	2	0.35	2	5	14	7	0

* Item point-biserials falling into this category were .04 and .20. While ideally, items should have a point-biserial of at least .25, these items had acceptable p-values (.26 and .86) and were retained to preserve adequate content coverage at the cluster level.

Table 4.2.5
Grade 6 - Frequency Distributions for Multiple-Choice p-values and Discrimination Indices by Content Area and Cluster

	Nitem	p-value						Discrimination					
		Median	p < 0.25	0.25 <= p < 0.50	0.50 <= p < 0.75	0.75 <= p < 0.90	p >= 0.90	Median	*pb < 0.20	0.20 <= pb < 0.30	0.30 <= pb < 0.40	0.40 <= pb < 0.50	pb >= 0.50
LAL	**20**	**0.71**	**0**	**2**	**11**	**7**	**0**	**0.34**	**1**	**6**	**10**	**3**	0
Interpreting Text	11	0.70	0	1	7	3	0	0.35	0	2	7	2	0
Analyzing Text	9	0.74	0	1	4	4	0	0.29	1	4	3	1	0
Math	**30**	**0.60**	**0**	**6**	**21**	**3**	**0**	**0.40**	**1**	**3**	**10**	**12**	**4**
Number and Numerical Operation	9	0.63	0	2	5	2	0	0.37	1	1	3	4	0
Geometry and Measurement	7	0.69	0	0	6	1	0	0.43	0	0	2	4	1
Patterns and Algebra	7	0.56	0	2	5	0	0	0.38	0	1	3	2	1
Data Analysis, Probability, and Discrete Mathematics	7	0.56	0	2	5	0	0	0.40	0	1	2	2	2
Problem Solving	28	0.60	0	5	21	2	0	0.40	1	3	9	11	4

* Item point-biserials falling into this category were .06 and .10. While ideally, items should have a point-biserial of at least .25, these items had acceptable p-values (.43 and .32) and were retained to preserve adequate content coverage at the cluster level.

Table 4.2.6
Grade 7 - Frequency Distributions for Multiple-Choice p-values and Discrimination Indices by Content Area and Cluster

	Nitem	p-value						Discrimination					
		Median	p < 0.25	0.25 <= p < 0.50	0.50 <= p < 0.75	0.75 <= p < 0.90	p >= 0.90	Median	*pb < 0.20	0.20 <= pb < 0.30	0.30 <= pb < 0.40	0.40 <= pb < 0.50	pb >= 0.50
LAL	**20**	**0.72**	**0**	**0**	**11**	**8**	**1**	**0.36**	**0**	**7**	**7**	**5**	**1**
Interpreting Text	12	0.70	0	0	8	4	0	0.34	0	5	4	3	0
Analyzing Text	8	0.82	0	0	3	4	1	0.38	0	2	3	2	1
Math	**30**	**0.47**	**2**	**16**	**12**	**0**	**0**	**0.33**	**3**	**4**	**15**	**4**	**4**
Number and Numerical Operation	7	0.57	0	3	4	0	0	0.35	0	1	4	2	0
Geometry and Measurement	9	0.42	1	4	4	0	0	0.33	1	1	5	1	1
Patterns and Algebra	7	0.38	1	5	1	0	0	0.35	1	1	3	1	1
Data Analysis, Probability, and Discrete Mathematics	7	0.47	0	4	3	0	0	0.31	1	1	3	0	2
Problem Solving	28	0.46	2	15	11	0	0	0.34	3	4	13	4	4

* Item point-biserials falling into this category ranged from -.11 to .20. While ideally, items should have a point-biserial of at least .25, these items generally had acceptable p-values (range .15 to .65) and were retained to preserve adequate content coverage at the cluster level. It should be noted that the scoring key for the one item with a negative point-biserial was checked and verified to be correct.

Table 4.2.7
Number of Dichotomous Items Where at Least One Distractor had a p-value or Point-Biserial Higher than or Equal to that of the Keyed Option

Test	Grade	N-item	p-value*	point-biserial*
LAL	5	20	0	0
	6	20	2	0
	7	20	0	0
Math	5	30	2	0
	6	30	2	0
	7	30	5	1

* The p-value and point-biserial in this table are calculated in the same way as for a correct answer, except in this case the distractor is used instead of the correct answer.

4.3 Speededness

The consequence of time limits on examinee's scores is called speededness. An exam is "speeded" to the degree that those taking it score lower than they would have if the test had not been timed. Most speededness statistics are based on the number of items that were not attempted by students. In each separately timed subsection of a test, if a student does not attempt the last item on the test, it would be assumed that the student may not have reached the item because they ran out of time.

The NJ ASK was not designed to be a speeded test, but rather a power test. That is, all students are expected to have ample time to finish all items and prompts. Because the tests were administered in three days, with multiple sessions each day, students were assumed to have enough time to complete the test. For example, a single writing prompt and a passage with 10 MC and 2 OE items were administered on the first day of testing for all grade levels. It included in two sessions a 25- (grade 5) or 45- (grades 6 and 7) minute session for the writing prompt and a 45-minute session for the reading test. Another passage and 10 MC and 2 OE items were administered on the second day. It included a 45-minute session. The mathematics tests were administered the third day. They included an 89-minute session with a 5-minute silent break after about 45 minutes.

Table 4.3.1 presents the percentage of students omitting the last MC item in each test session. For the LAL test, less than one percent of students omitted the last MC item in each session, while in the mathematics test, less than four percent of students omitted the last MC item.

Table 4.3.1
Percentage of Students Omitting the Last MC Item in Each Test Session

	Session	Item Location	Grade 5	Grade 6	Grade 7
Day One	LAL 2	LAL MC10	0.55%	0.79%	0.68%
Day Two	LAL 3	LAL MC20	0.38%	0.45%	0.44%
Day Three	Math	Math MC 29	1.04%	1.99%	2.66%
		Math MC 30	1.82%	1.97%	3.90%

4.4 Item Bias Statistics

Differential Item Functioning (DIF) was examined with the Mantel-Haenszel (1959)[10] procedure for the MC items and Winsteps (v3.60, Linacre, 2006)[11] for the OE items. Results are summarized in Table 4.4.1. The Mantel-Haenszel (MH) method is a non-parametric approach to DIF. In the MH procedure, total raw scores are held constant while the odds ratio is estimated. In testing practice, the odds ratio is generally converted to the delta metric and the ETS categorization is applied to flag the significance of DIF effects (Dorans & Holland, 1993)[12].

The critical values of the ETS categorizations are 1.00 and 1.50 on the delta scale for categories of A, B, and C. Specifically, if the absolute value of delta is smaller than 1.00 then the item is categorized as "A." If the absolute value of delta is larger than or equal to 1.50, then the item is classified as "C." Otherwise, items are categorized as "B." In both categories of A and C, statistical significance is set at the 5% level for a single item.

DIF detection with Winsteps is a Rasch-model-based approach. According to Linacre, Winsteps DIF detection contains the following steps:

- A joint run with all persons and all items is used to produce anchor values, i.e., ability and rating (or partial credit) scale structure.
- A subgroup run (reference group) with person abilities (or partial credit) scale structure anchored is used to produce group *R* item difficulties (D_R).
- Another subgroup run (focal group) with person abilities (or partial credit) scale structure anchored is used to produce group *F* item difficulties (D_F).
- DIF contrast ($D_F - D_R$) is obtained by taking the different DIF measures of the two-subgroup runs.
- A t-test statistic gives significance values as a unit normal deviate.

Rasch and Mantel-Haenszel procedures for DIF are equivalent under certain conditions (Linacre & Wright, 1989[13]; Schulz, Perlman, Rice, & Wright, 1996[14]). Similar to ETS classifications, the DIF output yielded by Winsteps is classified as negligible (A), slight to moderate (B), or moderate to severe (C). The classification criterion is as follows:

[10] Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of National Cancer Institute*, 22, 719-748.
[11] Linacre, M. (2006). Winsteps Rasch Measurement 3.60 [Computer software], Chicago, IL.
[12] Dorans, N. J. & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.
[13] Linacre, J. M. & Wright, B. D. (1989). The equivalence of Rasch, PROX, and Mantel-Haenszel. Transactions of the Rasch Measurement SIG, 3-2, 1-3.
[14] Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures for assessing Differential Item Functioning. In G. Engelhard & M. Wilson (Ed.), Objective measurement: Theory into practice (Vol 3). NJ: Ablex.

If a t-value is smaller than 2.58 or DIF contrast is smaller than 0.45 logits, then the item is flagged as "A." If a t-value is larger than 2.58 and the DIF contrast is larger than 0.65 logits, then the item is flagged as "C." Other items are flagged as "B." This categorization seems to be slightly more conservative than the ETS categorizations using the MH odds ratio (Liu & Mix, 2006)[15].

Table 4.4.1
Frequency Distribution of DIF Categories by Item Type and Grade Level

	Grade	DIF Category**	Dichotomous*			Open-ended*		
			M/F***	W/B***	W/H***	M/F	W/B	W/H
LAL	5	A	19	19	18	3	5	5
		B	1	1	2	2	0	0
		C	0	0	0	0	0	0
	6	A	19	17	19	5	5	5
		B	0	3	1	0	0	0
		C	1	0	0	0	0	0
	7	A	18	16	16	5	5	5
		B	2	4	3	0	0	0
		C	0	0	1	0	0	0
Math	5	A	29	30	29	3	3	3
		B	1	0	1	0	0	0
		C	0	0	0	0	0	0
	6	A	28	30	29	3	3	3
		B	2	0	1	0	0	0
		C	0	0	0	0	0	0
	7	A	30	29	30	3	3	3
		B	0	1	0	0	0	0
		C	0	0	0	0	0	0

* The Mantel-Haenszel procedure is applied for the MC items and Winsteps for the OE items.
** DIF categories: A, negligible; B, slight to moderate; and C, moderate to severe.
***DIF contrast groups: M/F, Male versus Female; W/B, White versus Black; and W/H, White versus Hispanic.

[15] Liu, Y. & Mix, D. (2006). Categorizations of Winsteps DIF output. Paper presented at the American Educational Research Association conference, San Francisco, CA.

4.5 Summary

The item analyses provided above generally show the NJ ASK to have sound psychometrics properties. For example, the distribution of p-values shows items to be measuring achievement across a broad range of difficulty. Also, item discrimination values show that most items are appropriately correlated with the total test score, and thus contribute to distinguishing between lower performing and higher performing students. In addition to the above, very few students omitted items from the tests. The low percentage of students omitting multiple-choice provides evidence that the test is a power test of the students' skills, and not their ability to pace themselves through a timed assessment. Finally, item bias statistics showed few NJ ASK items to demonstrate DIF. Items attaining the third category of DIF (i.e., C) will, if used for the operational NJ ASK assessment in 2007, be evaluated by a Riverside bias review committee and an independent sensitivity committee designated by the NJDOE.

PART 5: TEST ADMINISTRATION

5.1 Introduction

Great care is taken to assure standard administration of the NJ ASK. Close attention to detail is necessary to ensure that a student taking the test in one location has an equal opportunity to succeed as a student at another location. Information about the administration of NJ ASK is available in the *New Jersey Assessment of Skills & Knowledge Spring 2006 Test Coordinator Manual Grades 5–7***.** That information will not be fully replicated here, but the following elements are of importance to this technical report.

5.2 Determining Students for Whom a School Is Accountable

State regulations require that all students be included in the statewide assessment program and assessed annually. This includes limited English proficient (LEP) students and students with disabilities. Beginning in school year 2001–2002, students with disabilities were administered the Alternative Proficiency Assessment (APA) statewide.

All public schools, including those without assessed grades, are counted into the state's accountability system. All schools without assessed grades are counted as one unit with their respective receiving schools. This helps ensure closer vertical alignment of instructional services. In addition, special education students served in proprietary schools are counted in the sending schools' accountability results, which ensure that placement decisions are reviewed closely at the school and district level for optimum student academic performance.

To guarantee compliance with State regulatory requirements (N.J.A.C. 6A:8-4.4), a full academic year is defined as the term that begins on July 1 and ends on or about June 30. This date was established to accommodate the start of the district/school fiscal year and the allowance of academic programs and services offered to students prior to September. Any student enrolling in a school or district for the first time after July 1, up to the test administration date, is not considered to be enrolled for a full academic year. However, for making AYP determinations, a full academic year begins on July 1 and extends to the test administration date.

New Jersey does not include in the accountability system the results of any student enrolled less than one full academic year in a school for school accountability, or in a district for district accountability. This does not exclude from a district's accountability the results of those students who transfer from one school to another within a district.

5.3 Accommodations and Modifications

Testing accommodations and modifications allowed for NJ ASK grades 5–7 are extensive. Special education students must be tested using the modified testing procedures specified in each student's Individual Education Plan (IEP) and must be aligned to the Acceptable Accommodations or Modifications guidelines which are located in Appendix A of the *NJ ASK 2006 TCM Grades 5–7*. Students who are eligible under Section 504 of the Rehabilitation Act of 1973 must be tested using modified testing procedures that are specified in the student's 504 accommodation plan and must be aligned to the Acceptable Accommodations or Modifications guidelines which are located in Appendix A of the *NJ ASK 2006 TCM Grades 5–7*.

Accommodations are also allowed for Limited English Proficient (LEP) students taking the NJ ASK. Limited English Proficient accommodations do not apply to former LEP students, but only to students currently participating in a language assistance program. Accommodations include additional time, translation of the test directions into the student's native language, and use of a bilingual translation dictionary. Electronic dictionaries and Franklin translators are not permitted. Also, read-aloud directions for LEP students must be modified to reflect the accommodations listed above. It should be noted that LEP students in attendance for less than one academic year are exempt from the LAL assessment, but must take the mathematics assessment and should be enrolled in a language assistance program.

5.4 Testing Exemptions

The April 2006 test administration of the NJ ASK was for all students in grades 5, 6, and 7 in New Jersey. Even students attending out-of-district placements were required to take the NJ ASK. Only out-of-district placements that were approved testing sites could administer the NJ ASK. If a student attended a placement that was not an approved testing site, the student had to return to the home district to be tested or the home district had to send an examiner to the out-of-district placement to administer the test to the student. Test results of students attending out-of-district placements were aggregated into the school and district results of the students' home district. Out-of-district placements included the following:

1. Private schools for the handicapped (in-state or out-of-state)
2. Special services school districts
3. Educational services or joint commissions
4. District-paid placements in regular, alternative, or shared-time public high schools outside of the student's home district (in-state or out-of-state)

5.5 Administration of Tests

Detailed instructions for administering the NJ ASK were provided in the *New Jersey Assessment of Skills & Knowledge Spring 2006 Test Coordinator Manual Grades 5–7***.** The NJ ASK for grades 5–7 was administered on three consecutive days according to the following schedule:

- April 4, 2006 Regular Testing—LAL – Day 1, Sections 1 and 2
- April 5, 2006 Regular Testing—LAL – Day 2, Section 3
- April 6, 2006 Regular Testing—Mathematics
- April 10–13, 2006 Make-up Testing—Flexible
- April 17–20, 2006 Make-up Testing—Additional make-up time for districts that had a break the week of April 10. All testing had to be completed by April 20, 2006.

Testing was not to be scheduled immediately after an athletic event or an assembly. All test schedules were checked with the appropriate school officials to ensure that other school activities did not interfere with the test administration. Other test administration procedures included:

- All testing had to be scheduled in the morning. Exceptions included homebound and bedside students, as well as students attending out-of-district placements who were tested at that placement by staff from the student's home district.
- The district and school test coordinators (DTCs/STCs) were responsible for scheduling times and places for regular and make-up testing and for ensuring that all testing was completed according to the procedures and schedule described in the *Test Coordinator Manual* and in the *Examiner Manual*.
- Students who were required to test but were absent for the regular test administration had to be tested on the make-up dates.
- Students whose answer folders were voided during testing were considered to have attempted the test section. They were not allowed to retake or resume taking the voided test section during the make-up.
- Students who began a section of the test and did not complete it during the specified testing time were not allowed to complete the test section during the make-up period or any other time unless additional time was specified in their IEP or 504 plan.

Under the direction of the district test coordinator, the school test coordinator had to prepare a student roster for each examiner. Each roster contained the names of the students whom each examiner would supervise during testing, the names of the proctors assisting the examiner, and the room number. The student rosters were distributed to examiners before testing to allow examiners to prepare for the test administration. Examiners returned the student rosters to the school test coordinator immediately after

testing was completed. The school test coordinators kept a copy of each roster and returned the originals to the district test coordinator. The use of student rosters aided test administrators in remaining organized during test administration.

5.6 Test Security Procedures

The NJ ASK test booklets and its contents were treated as secure materials. Detailed procedures for maintaining the security of test materials while they were in the districts were outlined in the *New Jersey Assessment of Skills & Knowledge Spring 2006 Test Coordinator Manual Grades 5–7*. Each district is responsible for guaranteeing the security of the test materials. Examiners, proctors, and other school personnel were prohibited from copying, reading, discussing, or disclosing any test items before, during, or after test administration. When not being used during a test period, test materials were stored in a secure, locked location that was accessible only to individuals whose access was authorized by the school test coordinator. Inventory forms tracked test materials as they moved from one location to another in districts.

As part of the test development procedures, "breach" test forms and examiner manuals were prepared in the event of a security breach. If the NJDOE identified a security breach during the test administration window, the subcontractor immediately removed the NJ ASK test materials from the involved district or school. The test booklets for the content area affected were coded with a void code indicating a security breach. If the NJDOE determined that there was enough time for testing, the breach forms were delivered to the district and the test was administered to the affected students in the content area impacted by the security breach. For students re-tested during the test administration window, scores were reported based on the breach form. If a security breach was identified after the testing window, the impacted test booklets were coded with a security breach void code and no test results were reported for that content area. However, students received a score for the content area not impacted by the security breach.

5.7 Conclusion

As with any standardized accountability instrument, security is a pressing concern. The information outlined in this chapter and detailed in other cited documents indicates the intense attention paid to the security of the NJ ASK. At every stage detailed attention is devoted to preventing unauthorized access to the questions and to preventing retention by individuals of inappropriate records.

PART 6: SCORING

6.1 Introduction

Acceptable test construction and appropriate administrative procedures would be to no advantage if the scoring of test items were careless, inconsistent, inappropriate to New Jersey's standards, or otherwise ineffective. The following sections are intended to give more detail on the scoring process.

6.2 Multiple-Choice Items

Answer keys approved by the New Jersey Department of Education (NJDOE) were used to score the multiple-choice items after the responses had been scanned. Each item had a specific key (i.e., A, B, C, or D), which had been supplied and verified by the appropriate content specialists. All correct answers were assigned the value of "1" while incorrect answers were assigned the value of "0." At no time in this process was the original scanned answer overwritten, in case the key was determined to be incorrect during the post-scoring quality assurance check. After scoring was completed, simple item statistics were provided to the appropriate RPC content specialist to ensure that the correct keys were being applied. If a key changed, then the process was repeated until the scoring file was correct. The following data were reviewed for quality assurance:

- percent of students getting the question correct;
- correlation of the item to the test as a whole;
- correlation of each possible response option to the test as a whole;
- percentage of students choosing each response option (A, B, C, D or X-omits); and
- flags for items with high difficulty or low correlations.

6.3 Open-Ended Items and Writing Tasks

The 2006 NJ ASK open-ended questions and writing responses at grades 5, 6, and 7 required hand scoring by Measurement Incorporated (MI) personnel. The processes of selecting and training scorers, reading and scoring papers, and monitoring scoring are described below in detail.

6.4 Scoring Personnel

Because MI has been conducting the handscoring of writing and open-ended items for many years, the company already has available a large pool of qualified, experienced readers, whom they invite to return when projects are pending. MI informs them that a project is pending and invites them to return. In addition, MI routinely maintains supervisors' evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. MI employed many experienced readers for scoring the 2006 NJ ASK and recruited new scorers as well.

MI procedures for selecting new readers are very thorough. After advertising in local newspapers, with a job service, and elsewhere, and receiving applications, staff in MI's human resources department reviews the applications and schedules interviews for qualified applicants. Qualified applicants are those with a four-year college degree in English, language arts, education, mathematics, science, or a related field. Each qualified applicant must pass an interview by experienced MI staff, write an acceptable essay, and receive good recommendations from references. MI then reviews all the information about each applicant before offering employment.

6.5 Range Finding and Development of Scoring Guides

Range finding meetings were conducted to establish "true" scores for a representative sample of papers. Between 100 and 220 sample papers per task were chosen from the current test administration by MI leadership personnel. For items using specific rubrics, the rubrics were discussed and refined. The sample responses brought to the range finding meetings were selected from a broad representation of New Jersey school districts in order to ensure that the sample is representative of overall student performance. The range finding committees consisted of NJDOE content specialists, New Jersey teacher representatives, MI management personnel, as well as the scoring director responsible for each content area and a Riverside content specialist for mathematics.

Each open-ended mathematics item for NJ ASK 5–7 was scored using an item-specific 4-point rubric (score points possible are 0, 1, 2, or 3). Creation of an item-specific rubric for mathematics was accomplished in the following manner:

- All rubrics represented a refinement of the NJDOE generic 4-point rubric which can be found at the following link: http://www.njpep.org/assessment/TestSpecs/MathTestSpec/GEPAMath/HolisticScoringGuide.html
- Each rubric for the mathematics open-ended items (9 rubrics in all, 3 items per grade) was then made item specific, incorporating specific knowledge, skills, and levels of student performance required by the item. This work was performed by Measurement Incorporated.
- The rubrics were further refined in range finding. At range finding, representatives of Measurement Incorporated, Riverside, New Jersey educators, and NJDOE content specialists scored 100 student papers, per item, using the item

specific rubrics created by Measurement Incorporated. Further refinement included incorporating into the rubric specific knowledge, skills, and problem solving strategies and the specific score point they would define.

- Finally, as student papers were scored, scoring questions were referred to the NJDOE content specialist for final adjudication.

After range finding, MI management and the scoring directors developed training materials consisting of an anchor set (examples of responses for each score point) and training/qualifying sets (practice papers) for each task using the responses scored at range finding. Anchor sets usually consisted of three or more annotated examples of each score point in score point order. Training/qualifying sets consisted of clearly anchored papers in random score point order.

6.6 Project Leads

In selecting team leaders, MI's management staff and scoring directors reviewed the files of all the returning staff. Individuals who are experienced team leaders with a record of good performance on previous projects are considered. Also considered are readers who have been recommended for promotion to the team leader position.

It's important to note that MI is an equal opportunity employer that actively recruits minority staff. Historically, temporary staff on major projects average about 70% female, 30% male, 76% Caucasian, and 24% minority. MI does not discriminate against any employee or applicant for employment with respect to hire, tenure, terms, conditions, or privileges of employment, or any matter directly or indirectly related to employment, because of race, color, religion, sex, age, handicap, national origin, ancestry, veteran status, or sexual orientation.

6.7 Training Team Leaders

After the anchor papers, training, and qualifying papers have been identified and finalized, team leader training is conducted by the scoring director for each task, a process which typically takes up to four days depending on the content. Procedures are similar to those for training scorers but are more comprehensive, dealing with resolution of discrepant scores, identification of non-scorable responses, unusual prompt treatment, alert situation responses (e.g., child-in-danger), and other duties performed only by team leaders. Team leaders took careful notes on the training papers in preparation for discussion with the scorers. Also, the scoring directors counseled team leaders on application of the rubric and training techniques. Effective scorer training relies to a great extent on having knowledgeable, flexible team leaders. Team leaders assisted in training scorers in discussions of training sets, and were responsible for distributing, collecting, and accounting for training packets and sample papers during each scoring session. During scoring, team leaders responded to questions, spot-checked scorer packets, and counseled scorers having difficulty with the criteria.

Team leaders also administered the quality control validity sets, monitored the scoring patterns of each scorer throughout the project, conducted retraining as necessary, performed some resolution readings, and maintained a professional working environment. Team leaders worked approximately 7.75 hours per day, excluding breaks.

6.8 Evaluators

Levels of staffing for scoring of the 2006 NJ ASK are listed in Table 6.8.1. The table shows the numbers of scorers, team leaders and scoring directors at each grade level who participated in scoring. Every scorer had a minimum qualification of a 4-year degree, and many had higher degrees.

Table 6.8.1
Number of Scorers, Training Leaders, and Scoring Directors at Each Grade

	2006 NJ ASK		
Grade	Scorers	Team Leaders	Scoring Directors
5	281	37	7
6	246	28	7
7	273	29	8

6.9 Scoring Procedures

All responses requiring hand scoring were scored by MI. All open-ended items for mathematics were scored on a scale of 0 to 3 while LAL was scored on a scale of 0 to 4. Writing was scored on a scale of 0 to 5 for grade 5 and 0 to 6 for grades 6 and 7. Note that for grades 6 and 7, the writing scores were doubled in data analyses and score reporting. Each response was scored at least once. Ten percent of the responses were randomly assigned to a second reader in order to measure inter-reader reliability (see Part 12 of this Technical Report). Scorers scored each student response without seeing any student biographical or demographic information. Evaluators viewed each response and either assigned a score or a condition code.

Training Scorers and Qualifying

All scorers were trained using the rubrics, anchor papers, training papers, and qualifying papers selected during the range finding meetings and approved by the NJDOE. Scorers were assigned to a scoring group consisting of one team leader and 10 to 12 scorers. Each scorer was assigned an individual number for easy identification of his or her scoring work throughout the scoring session.

After contracts and nondisclosure forms were signed, training began. Scorer training followed the same format as team leader training. The scoring director introduced the set

of anchor papers and thoroughly discussed each score point. This presentation was followed by practice scoring on the training sets. Scorers broke into teams to discuss the papers in the training sets. This arrangement gave scorers an opportunity to discuss any possible points of confusion or problems in understanding the criteria in a small group setting.

Team leaders collected the monitor sheets after the scoring of each training set, and recorded results on a customized log which was examined by the scoring director to determine which papers might give scorers difficulty. The scoring director also "floated" from team to team, listening to the team leaders' explanations and adding additional information when necessary. If a particular paper or type of paper seemed to be causing difficulty across teams, the problem was discussed with the room at large to ensure that everyone heard the same explanation.

For qualifying, scorers had to demonstrate their ability to score accurately by attaining 90% adjacent agreement (within one point) percentage on the qualifying sets before they read packets of actual papers. Any reader unable to meet the standards set by the NJDOE was dismissed. All scorers understood this stipulation when they were hired.

Training was carefully orchestrated so that scorers understood how to apply the rubric in scoring the papers, how to reference the scoring guide, develop the flexibility needed to deal with a variety of responses, and retain the consistency needed to score all papers accurately. In addition to completing all of the initial training and qualifying, a significant amount of time was allotted for demonstrations of paper flow, explanations of "alerts" and "flagging," and instructions about other procedures which were necessary for the conduct of a smooth project. Scorers generally worked seven hours per day, excluding breaks.

Monitoring for Quality Assurance

MI constantly monitors the quality of each scorer's work throughout every project. Methods used to monitor scorers' scoring habits in scoring NJ ASK included the use of Daily Reader Status Reports.

Each student writing sample was scored holistically by readers using the Registered Holistic Scoring Method. Previously, a different reader from another team read identified packets a second time. At no time did readers see the other's scores. After the scores from each day's work are entered, MI's data application calculated the results and generated a status report. These reports showed the total number of papers read and the percentage agreement of each reader, both perfect and adjacent, for the second-read packets. The reports also showed score point distributions. Scoring directors examined the reports and used the information to determine the need for retraining of individual readers or the group as a whole. It could easily be determined if a reader was consistently scoring "too high" or "too low," as well as the specific score points with which they may be having difficulty. The Daily Reader Status Reports showed not only the current daily totals for each scorer, but also the project-to-date totals.

Retraining was an ongoing process once scoring began. Daily monitoring of completed packets and analysis of agreement rates provided by the Daily Reader Status Reports and

validity packets alerted team leaders and management personnel to individual retraining needs. If it became apparent that a whole team or a whole group was having difficulty with a particular type of response, large group training sessions were conducted. Standard retraining procedures included room-wide discussions led by the scoring director, team discussions conducted by team leaders, spot-checking of individual scorers by team leaders, and discussions between team leaders and individual scorers.

Scorers were dismissed when, in the opinion of the scoring director and the project director, they had been counseled, retrained, and given every reasonable opportunity to improve but were still performing below the acceptable standard.

6.10 Conclusion

Based upon the above procedures, scoring directors and team leaders carefully monitored each scorer's individual performance and the performance of each scoring group as a whole. This allowed any potential scoring drift to be identified and corrected through further training. Scorers were monitored for scoring accuracy and consistency daily. Throughout the scoring period, the reliability of each evaluator was checked through read-behinds and validity packets.

PART 7: SCALING AND EQUATING

7.1 Introduction

This section details the scaling and equating procedures used for the 2006 NJ ASK, including how the scoring tables were created. Scaling and equating was accomplished using Winsteps (Linacre, 2006)[16]. Winsteps is designed to produce a single scale by jointly analyzing data resulting from students' responses to both multiple-choice and open-ended items. In Winsteps, items are calibrated based on item response theory (IRT), using the one-parameter logistic model for multiple-choice items (Rasch, 1960; Wright and Stone, 1979)[17] and the partial credit model (Masters, 1982)[18] for open-ended items. Winsteps was also used to equate the scales developed by two calibrations through a common-item procedure.

7.2 Item Response Theory

Hambleton (1989)[19] writes, "In its most common and popular form, item response theory postulates that (a) underlying examinee performance on a test is a single ability or trait, and (b) the relationship between examinee performance on each item and the ability measured by the test can be described by a monotonically increasing function" (p. 148). Central to IRT is the application of mathematical functions designed to model or describe the relationship between performance on each item on a test and the amount of the ability or trait the item is measuring. These functions are commonly called item-characteristic curve (ICCs) (Hambleton & Swaminathan, 1985)[20].

Item-characteristic curves represent the relationship between the amount of the trait or ability an item measures (in IRT ability level is commonly denoted by θ) and the probability of a correct response to the item (denoted $P_i(\theta)$). When items are scored dichotomously (i.e., scored right or wrong), Hambleton (1989) suggests that the probability be interpreted "as the probability associated with a randomly selected examinee at ability level θ, answering item i correctly" (p. 153).

[16] Linacre, J. M. (2006). A *User's Guide to WINSTEPS MINISTE Rasch-Model Computer Programs*. Chicago

[17] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.

[18] Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

[19] Hambleton, R. K (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: American Council on Education.

[20] Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer.

One of the most popular IRT models is the Rasch, or one-parameter logistic, model (Rasch, 1960; Wright & Stone, 1979). Rasch (1960) takes his readers step-by-step through his thought processes and the circumstances under which this important model was derived. Simply stated, the Rasch model (a mathematical function that would produce an ICC) is:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

where $P_i(\theta)$ = the probability of a correct response to item i given θ,

θ = ability level,

b_i = the difficulty parameter for item i.

Two important points to note regarding the Rasch model are (a) each model has only one person parameter (i.e., θ) to estimate and one item parameter (i.e., b_i) to estimate, and (b) the primary "goal" of each model is to describe the relationship between the amount of a trait or ability (θ) and "success" on an item. Also, IRT requires several assumptions to be made for the derivation of models and the estimation of parameters within the models. One important IRT assumption is unidimensionality (Hambleton, 1989; Hambleton & Swaminathan, 1985). With respect to unidimensionality, IRT assumes that only one trait or ability underlies test performance.

Finally, Masters (1982) developed an extension of the Rasch model to handle polychotomous items, or items that allow for responses that give less than full credit (e.g., open-ended items). Masters' extension of the Rasch model is called the partial credit model. As noted above, all NJ ASK item calibrations used the Rasch model for multiple-choice items (dichotomous items) and the partial credit model for open-ended items (polychotomous items).

7.3 Scaling and Equating

Scaling and equating was accomplished using Winsteps. The program uses IRT to produce a single scale by jointly analyzing data resulting from students' responses to both multiple-choice and open-ended items. In Winsteps, a Joint Maximum Likelihood Estimation (JMLE) method is implemented. Equating was conducted with anchored calibrations. That is, a first run was executed using all students where all items and person parameters were calibrated without constraint. Next, a second run was executed fixing the item parameters of the common items between the "new" test (e.g., the Braille test form) and the "original" test (i.e., the operational test form for regular students).

The 2006 items were generally calibrated before item response data were available for all students for whom reports were to be generated. For most grades, the early return data used for standard setting was also used for item calibrations. For LAL grades 5 and 6, Rasch parameters were calibrated based on a more complete data set. Since no Grade 5 students obtained a maximum possible score on one open-ended item in LAL, a student who obtained the highest total raw score was assigned a perfect score to this open-ended item for the purpose of calibration. This allowed Winsteps to produce an estimate for every possible number correct, raw score total. Table 7.3.1 displays the number of students used for the calibrations by grade and content area in 2006.

Table 7.3.1
Number of Students Used for Calibrations
by Grade and Content Area for NJ ASK in 2006

	Number of Students	
Grade	LAL*	Mathematics
5	54,698	14,185
6	49,621	14,208
7	14,182	14,199

* Item calibrations based upon standard setting sample, except LAL for grades 5 and 6 where more student data was used to obtain each score point.

Table 7.3.2 summarizes Infit and Outfit statistics for the 2006 NJ ASK tests. The Infit statistic is more sensitive to unexpected behavior affecting responses near an examinee's ability level while the Outfit statistic is more sensitive to unexpected behavior by examinees far from their ability level (see Winsteps Manual, pp. 201–202).

Infit and Outfit can be expressed as a mean square (MNSQ) statistic or on a standardized metric (ZSTD). MNSQ values are more oriented toward practical significance, while Z values are more oriented toward statistical significance. As a rule of thumb (pp. 201–202), an item is considered productive of measurement if MNSQ is within 0.50–1.5. For the 2006 NJ ASK tests, few items had a MNSQ value outside of the range. Although no Infit MNSQ laid outside of the range 0.50–1.50, both the mathematics test and the LAL test for grade 6 had one item with an Outfit MNSQ larger than 1.50. These items are listed below:

- LAL Grade 6 item 21
- Math Grade 5 item 14
- Math Grade 6 item 12
- Math Grade 7 item 27

All items were multiple-choice (MC) items. Each item had a relatively low p-value (high difficulty) and/or low point-biserial.

Table 7.3.2
Summary for the Infit and Outfit Statistics
for the 2006 NJ ASK by Grade and Content Area

		MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
LAL 5	MEAN	0.00	0.01	0.99	-0.9	0.99	-0.7
	S.D.	1.77	0.00	0.08	9.0	0.15	8.8
	MAX.	4.51	0.01	1.13	9.9	1.26	9.9
	MIN.	-2.31	0.00	0.87	-9.9	0.63	-9.9
LAL 6	MEAN	0.00	0.01	1.00	-0.7	1.02	-0.4
	S.D.	1.50	0.00	0.11	8.5	0.19	8.6
	MAX.	3.01	0.01	1.29	9.9	1.53	9.9
	MIN.	-2.13	0.00	0.77	-9.9	0.71	-9.9
LAL 7	MEAN	0.00	0.01	0.99	-0.9	1.01	-0.1
	S.D.	1.36	0.00	0.12	8.7	0.22	9.0
	MAX.	2.77	0.01	1.17	9.9	1.47	9.9
	MIN.	-2.39	0.00	0.80	-9.9	0.63	-9.9
Math 5	MEAN	0.00	0.01	1.00	-0.3	1.00	-0.5
	S.D.	1.06	0.00	0.09	8.9	0.19	8.7
	MAX.	2.45	0.01	1.25	9.9	1.67	9.9
	MIN.	-1.98	0.00	0.87	-9.9	0.67	-9.9
Math 6	MEAN	0.00	0.01	1.00	0.4	1.00	0.4
	S.D.	0.75	0.00	0.10	8.7	0.19	8.3
	MAX.	1.45	0.01	1.31	9.9	1.73	9.9
	MIN.	-1.66	0.00	0.84	-9.9	0.68	-9.9
Math 7	MEAN	0.00	0.01	1.00	-0.1	1.04	0.2
	S.D.	0.82	0.00	0.11	9.1	0.25	8.8
	MAX.	1.77	0.01	1.35	9.9	2.23	9.9
	MIN.	-1.36	0.00	0.81	-9.9	0.76	-9.9

Equating Procedures

This section describes the equating procedures for scores from the Large Print, Braille, and Breach Forms of the 2006 NJ ASK. Large Print and Braille test forms were constructed by removing one or two items from the corresponding regular test forms. Items that were removed from the regular test forms are summarized in Table 7.3.3.

Table 7.3.3
Summary of Items that Were Removed from the Regular Assessment

	Grade	Large Print	Braille	Maximum Possible Score Remained
LAL	5	N/A	Writing Prompt	36
Math	5	Items #1 and #16	Items #1 and #16	37
	6	Item #13	Item #13	38
	7	Items #32 and #33	Items #32 and #33	35

Several assumptions had to be made in order to equate the scores of the Large Print and Braille tests to the scores of the regular test. First, it was assumed that the latent trait measured by these tests and the regular test was the same. Given the fact that the same items were used across the tests within each content area, with the exception of the removed items, it seemed reasonable to assume that changes to item format or item presentation would not greatly change the *overall* latent trait or construct measured by each assessment.

A second and stronger assumption, however, was that item parameters across the tests within each content area were identical. This of course is a very strong assumption considering the different item formats across the tests. However, this assumption was necessary because sample sizes for the Large Print and Braille tests were too small to get reliable parameter estimates. Because the first assumption noted above is reasonable, i.e., for each test the LAL assessment measures language arts and the mathematics assessment measures mathematics, the following steps for equating the Large Print and Braille tests to the regular test were used:

- Conduct an anchored item calibration. The above items were removed and the parameters and steps of the Large Print or Braille test items were fixed with the estimates resulting from the corresponding regular test items.
- Transform the theta metric to the scale score metric. Because the theta values obtained from the anchored calibration and those obtained from the regular test score calibration are on the same metric, the transformation functions applied to the regular test scores can be applied to the Large Print and Braille test scores.
- Create raw score to scale score look-up tables for each Large Print and Braille test. In cases where no raw score corresponds to the cut scale scores (200 for Proficient and 250 Advanced Proficient), the raw score point immediately below the cut score was assigned as the cut point scale score.

A total of 376 Grade 6 students were administered the Breach test forms. For mathematics, the Breach form was the same as the regular test form, except the items were reordered. For LAL, in addition to reordering the items, the writing prompt was replaced with a new prompt. Because the performance score distributions and results of the LAL Breach score equating were so similar to the operation forms, it was decided to apply to the Breach forms the same raw to scale score conversion table as used with the regular test forms. The similarity between the calibrations no doubt resulted from the fact that all items were the same on both forms except the writing prompt.

Producing the Scoring Tables

Total scores for NJ ASK 2006 were reported in scale scores with a range of 100–300. Note that scores of 100 and 300 were a theoretical floor and ceiling and may not actually have been observed for some grades and/or content areas. The scale score of 200 represents the cut point between Partially Proficient (PP) and Proficient (P) while the scale score of 250 represents the cut point between Proficient and Advanced Proficient (AP). The scale score ranges are as following:

Partially Proficient	100 to 199
Proficient	200 to 249
Advanced Proficient	250 to 300

NJ ASK 2006 scale scores are linearly related to the theta metric calibrated using Winsteps. Producing the scoring tables included in the following steps:

- decide cut score points on the raw score metric through standard setting;
- calibrate Rasch parameters with NJ ASK 2006 standard setting sample data (except LAL grades 5 and 6);
- find cut score points on the theta metric;
- calculate intercept and slope of theta-to-scale-score transformation function; and
- create raw score to scale score conversion tables.

As previously noted, for LAL grades 5 and 6, Rasch parameters were calibrated based on a more complete data set. Since no Grade 5 students obtained a maximum possible score on one open-ended item in LAL, a student who obtained the highest total raw score was assigned a perfect score to this open-ended item for the purpose of calibration. This allowed Winsteps to produce an estimate for every possible number correct, raw score point total.

Standard-setting procedures were described in Part 3 of this Technical Report and in greater detail in another document (*Standard setting for the New Jersey assessment of skills and knowledge for grades 5, 6, and 7*, Riverside Publishing, 2006). Cut scores established through the standard setting are summarized in Table 7.3.4.

Table 7.3.4
NJASK 2006 Cut Scores Established through the Standard Setting

		Language Arts Literacy		Mathematics	
		Proficient	Advanced Proficient	Proficient	Advanced Proficient
Grade 5	Cut Score	16	30	18	30
Grade 6	Cut Score	20	32	17	31
Grade 7	Cut Score	21	34	13	26

Linear transformations were applied to theta estimates and scale scores. The following formula was used to obtain the slopes and intercepts for the transformation functions:

$$sc(y) = \left[\frac{sc(y_2) - sc(y_1)}{\theta_2 - \theta_1}\right] y + \left\{(sc(y_1) - \left[\frac{sc(y_2) - sc(y_1)}{\theta_2 - \theta_1}\right]\theta_1\right\},$$

where θ_1 and θ_2 are person parameter estimates that correspond to the cut score points, $sc(y_1)$ and $sc(y_2)$ are scale score points.

The above formula was adopted from Kolen and Brennan (2004, p. 337[21]). In NJ ASK 2006, $sc(y_1)$ is 200 and $sc(y_2)$ is 250. Slopes and intercepts of the transformation functions are summarized in Table 7.3.4. A raw score to scale score look-up table for each test form is attached as Appendix 7-1.

Table 7.3.4
Summary of Slopes and Intercepts of Theta to Scale Score Transformation Functions by Grade Level and Content Area

Test	Grade	Proficient			Advanced Proficient			Slope	Intercept
		RS	Theta	SS	RS	Theta	SS		
LAL	5	16	-0.529	200	30	2.266	250	17.8891	209.4633
	6	20	0.025	200	32	1.675	250	30.3030	199.2424
	7	21	0.230	200	34	2.165	250	25.8398	194.0568
Math	5	18	0.099	200	30	1.553	250	34.3879	196.5956
	6	17	-0.096	200	31	1.700	250	27.8396	202.6726
	7	13	-0.579	200	26	0.990	250	31.8674	218.4512

[21] Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and practice*. NY: Springer.

PART 8: REPORTING

8.1 Introduction

Scores are reported in two cycles. Cycle I data are considered preliminary. For additional information about score reports that are very similar to the NJ ASK Cycle I and Cycle II reports, see the GEPA Score Interpretation Manual.[22]

8.2 Cycle I Reports

The Cycle I reports included the following, separate for each grade: Individual Student Report, Student Sticker, All Sections Roster, Summary of School Performance, and Summary of District Performance. Each Cycle I report is briefly described below.

Individual Student Report

The Individual Student Report (ISR) is a two-sided report, produced in alphabetical sequence for students within the school. The ISR for NJ ASK at grade 5 is shown in sample format as Figure 8.2.1 (front page). Two copies of this report are produced for every student tested, one for the student's permanent folder, and the other for the student's parent/guardian to be shared in a manner determined by the local district.

Test Date: APRIL 2006
Report Printed: 10/18/2006

New Jersey Assessment of Skills and Knowledge
Individual Student Report

NJ ASK

County:	88 ANY COUNTY	Date of Birth: 04/09/91	LEP: <	Answer Folder No: 40225
District:	7777 ANY DISTRICT	Sex: F	SE:	District/School ID No: 19555
School:	666 ANY SCHOOL	Out of District Placement:	LEP Exempt (LAL Only): E	Special Form:
		Out of Residence Placement: Y	Title I:	Grade: 5

Student Name: STUDENT A, HUI
ASK ID No: 1206201718

Content Area	Your Child's Scale Score	Proficiency Level
Mathematics	169	PARTIALLY PROFICIENT
Language Arts Literacy	EXEMPT FROM TAKING	

Partially Proficient / Scale Score BELOW 200
Proficient / Scale Score AT OR ABOVE 200 but BELOW 250
Advanced Proficient / Scale Score AT OR ABOVE 250

Mathematics

The Mathematics section assesses a student's abilities in the following clusters.

Cluster	Your Child's Points	Just Proficient Mean
Number & Numerical Operations	2.0 out of 10.0	4.8
Geometry & Measurement	2.0 out of 9.0	5.1
Patterns & Algebra	2.0 out of 10.0	3.9
Data Analysis, Probability & Discrete Mathematics	6.0 out of 10.0	4.2
Problem Solving	9.0 out of 37.0	16.8
Your Child's Total Points	**12.0 out of 39.0**	

Language Arts Literacy

The Language Arts Literacy section assesses a student's abilities in the following clusters.

Cluster	Your Child's Points	Just Proficient Mean
Writing		
Reading		
Working with Text		
Analyzing Text		
Your Child's Total Points		

[22] For the GEPA Manual, see: http://www.state.nj.us/njded/assessment/ms/gepa_score_interp_manual.pdf.

Figure 8.2.1: Individual Student Report

The scale scores in LAL and mathematics, along with cluster data, are provided on the front of the ISR (see Figure 8.2.1). Explanatory text about scale scores, proficiency levels, and cluster scores is provided on the back of the ISR.

The Just Proficient Mean is a statewide statistic comprised of the average or mean score attained on each cluster by all students by grade—i.e., general education students (GE), special education students (SE), and limited English proficient students (LEP)—with a scale score of 200—i.e., students who are "just proficient." Students whose NJ ASK test booklets were coded as "void" were excluded from these means, along with any students who took an alternate form (i.e., Braille, Large Print, or Breach).

Student Sticker

The Student Sticker is produced alphabetically, and one sticker for each student within the school is provided. It is a peel-off label designed to be easily attached to the student's permanent record. The scale sores in LAL and Mathematics are provided. Designations of the proficiency levels are printed next to the LAL and mathematics scale scores. Voids, where applicable, are noted.

All Sections Roster

The All Sections Roster (Figure 8.2.2) provides a convenient method for reviewing students' complete test results. The report displays student names in alphabetical order (last name first). Users of this report can quickly determine how a particular student performed in both content areas: Language Arts Literacy and mathematics.

TEST DATE: APRIL 2006
REPORT PRINTED: 10/17/2006

New Jersey Assessment of Skills and Knowledge
All Sections Roster
Grade 5

NJ ASK

COUNTY: 88 ANY COUNTY
DISTRICT: 7777 ANY DISTRICT
SCHOOL: 666 ANY SCHOOL

STUDENTS PROCESSED: 40
(EXCLUDES STUDENTS WHO TOOK BRAILLE, LARGE-PRINT, ALTERNATE OR SPECIAL EQUATED FORMS.)

STUDENT NAME/ ASK ID NUMBER	DOB	SEX	EC	LEP	SE	LEP EXEMPT (LAL Only)	T-I	ED	MI	OUT OF DIST	OUT OF RES	TID <1	TIG <1	MATHEMATICS SCALE SCORE	MATHEMATICS PROFICIENCY LEVEL	LANGUAGE ARTS LITERACY SCALE SCORE	LANGUAGE ARTS LITERACY PROFICIENCY LEVEL
STUDENT A, JAMES 1380282225	06/04/92	M												271	ADVANCED PROFICIENT	270	ADVANCED PROFICIENT
STUDENT B, REBECCA L 1380289434	01/11/91	F	W									Y	Y	266	ADVANCED PROFICIENT	275	ADVANCED PROFICIENT
STUDENT C, RADHAMES 1330027318	05/11/92	M												188	PARTIALLY PROFICIENT	217	PROFICIENT
STUDENT D, KIRYL 1380275530	01/28/92	M												V1		247	PROFICIENT
STUDENT E, VINCENT 1380038224	11/11/91	M	W											192	PARTIALLY PROFICIENT	191	PARTIALLY PROFICIENT
STUDENT F, ALFREDO 1330028117	09/09/91	M	H	F										185	PARTIALLY PROFICIENT	220	PROFICIENT
STUDENT G, DIVINITY 1380278331	02/06/91	F	W						Y					214	PROFICIENT	231	PROFICIENT
STUDENT H, GABRIELLA R 1380142321	06/17/91	F	W	1										193	PARTIALLY PROFICIENT	176	PARTIALLY PROFICIENT
STUDENT I, LAURA 1380282629	03/27/92	F	H	1										175	PARTIALLY PROFICIENT	203	PROFICIENT
STUDENT J, VIVIANN M 1380149530	01/23/90	F	H	F	I			Y						NOT PRESENT		115	PARTIALLY PROFICIENT

Figure 8.2.2: All Sections Roster

Following a student's identification information, the student's Scale Score and Proficiency Level (Partially Proficient, Proficient, or Advanced Proficient) are printed for each test section. If the student's test booklet was coded void, the reason code appears in this space.

Performance Summaries (District and School)

Figure 8.2.3 provides an example of the Summary of District Performance report for each content area (Language Arts Literacy and mathematics) and grade. The report is produced at the district level and provides preliminary aggregated data. Final aggregated data are produced in Cycle II reporting as there are inevitably a few changes made to data between Cycle I and Cycle II. Data are provided for total students, general education students, special education students, and limited English proficient students. The report includes the percent and number of students in each proficiency level for total students, general education students, special education students, and limited English proficient students tested. The report also provides Just Proficient Means as well as scale score means and cluster means for the aforementioned groups.

TEST DATE: APRIL 2006
REPORT PRINTED: 10/16/2006

New Jersey Assessment of Skills and Knowledge
Summary of District Performance - Mathematics
Grade 5

NJ ASK

COUNTY: 88 ANY COUNTY
DISTRICT: 7777 ANY DISTRICT

STUDENTS PROCESSED: 74 (EXCLUDES STUDENTS WHO TOOK BRAILLE, LARGE-PRINT, ALTERNATE OR SPECIAL EQUATED FORMS.)

PERFORMANCE DATA FOR STUDENTS WITH VALID SCALE SCORES [1]

DEMOGRAPHIC GROUP	Number Processed	Not Present	Voids	Valid Scale Scores	Partially Proficient Number	Partially Proficient Percent	Proficient Number	Proficient Percent	Advanced Proficient Number	Advanced Proficient Percent
Total Students [2]	74	6	2	66	41	62.1	20	30.3	5	7.6
General Education	55	0	2	53	28	52.8	20	37.7	5	9.4
Special Education	6	6	0	0	0	0.0	0	0.0	0	0.0
Limited English Proficient [3]	15	2	0	13	13	100.0	0	0.0	0	0.0
LEP Exempt (LAL Only)	N/A	--	--	--						
Students Coded Both SE and LEP:	2	2	0	0						

SCALE SCORE MEANS & CLUSTER MEANS FOR STUDENTS WITH VALID SCALE SCORES

	Scale Score Mean [4]	Number & Numerical Operations	Geometry & Measurement	Patterns & Algebra	Data Analysis, Probability & Discrete Mathematics	Problem Solving
Total Points Possible		10.0	9.0	10.0	10.0	37.0
Just Proficient Mean [5]		4.8	5.1	3.9	4.2	16.8
Total Students [2]	196.5	3.5	4.6	5.2	5.7	14.4
General Education	202.1	3.9	4.9	5.5	6.1	15.5
Special Education	0.0	0.0	0	0	0.0	0.0
Limited English Proficient [3]	174.0	2.0	3.2	4.1	4.4	9.6

[1] Percentages may not total 100 due to rounding.
[2] Students are included in Total Students only once, but they appear in all other categories that apply.
[3] Excludes students coded Former LEP.
[4] The range of scale scores is 100 to 300.
[5] Just Proficient Mean is the statewide raw score mean for students whose scale score is 200.

88-7777

Figure 8.2.3: Summary of District Performance

There is also a Summary of School Performance report for each content area and grade, which provides aggregated data for the school. This report includes data for total students, general education students, special education students, and limited English proficient students. The report format is the same as the Summary of District Performance.

8.3 Cycle II Reports

The Cycle II reports, produced separately for each grade, include the following: Performance by Demographic Group School, Performance by Demographic Group District, and Performance by Demographic Group Statewide. These Cycle II reports are briefly described below.

Performance by Demographic Groups (Statewide, District, and School)

The school, district, and statewide reports provide a complete analysis of student performance for each content area tested. These reports summarize student performance on the total test sections, disaggregated by population, gender, ethnicity, economic states and migrant status. An example of the statewide report is given in Figure 8.2.4.

TEST DATE: APRIL 2006
REPORT PRINTED: 07/31/2006

New Jersey Assessment of Skills and Knowledge
Performance By Demographic Group
Grade 5
Statewide

NJ ASK

Demographic Group	Mathematics								Language Arts Literacy							
					Performance Data for Students with Valid Scale Scores[2]								Performance Data for Students with Valid Scale Scores[2]			
	Number Enrolled[1]	Number Not Present	Number of Voids	Number of Valid Scale Scores	% Partially Proficient	% Proficient	% Advanced Proficient	Scale Score Mean	Number Enrolled[1]	Number Not Present	Number of Voids	Number of Valid Scale Scores	% Partially Proficient	% Proficient	% Advanced Proficient	Scale Score Mean
Total Students [3]	104,284	306	165	103,813	18.2	54.2	27.5	227.7	104,284	289	169	103,248	14.1	76.6	9.3	220.1
General Education	84,895	184	65	84,646	11.6	56.4	32.0	234.0	84,895	115	65	84,715	7.6	81.3	11.1	224.0
Special Education	16,282	117	96	16,069	46.1	45.9	8.0	200.3	16,273	78	103	16,092	41.8	56.9	1.3	202.8
Limited English Proficient[4]	3,295	7	5	3,283	54.7	39.4	6.0	194.6	2,717	97	2	2,618	57.3	42.0	0.7	196.4
LEP Exempt (LAL Only)	N/A	--	--	--	--	--	--	--	578	--	--	--	--	--	--	--
Gender [5]																
Female	50,639	123	60	50,456	17.1	56.1	26.8	227.8	50,639	117	58	50,180	11.7	77.2	11.1	221.9
Male	53,623	174	105	53,344	19.3	52.5	28.2	227.5	53,623	163	110	53,056	16.3	76.0	7.7	218.3
Migrant Status																
Migrant	37	1	2	34	35.3	55.9	8.8	208.2	37	1	1	28	35.7	60.7	3.6	206.0
Non-Migrant	104,247	305	163	103,779	18.2	54.2	27.5	227.7	104,247	288	168	103,220	14.1	76.6	9.3	220.1
Ethnicity [6]																
White	58,223	110	92	58,021	10.1	55.9	34.0	235.8	58,223	89	87	57,960	6.7	81.3	11.9	225.3
Black	18,169	113	41	18,015	37.9	52.3	9.8	206.7	18,169	66	47	18,042	30.4	67.4	2.3	208.1
Asian	7,654	11	3	7,640	6.7	41.7	51.6	248.4	7,654	18	3	7,540	5.6	73.5	20.9	230.4
Pacific Islander	309	1	0	308	7.1	48.7	44.2	242.8	309	0	0	309	6.1	77.0	16.8	228.2
Hispanic	18,808	60	26	18,722	29.5	56.1	14.4	213.9	18,808	104	27	18,303	24.9	72.0	3.1	210.9
American Indian/Alaskan Native	107	0	2	105	24.8	63.8	11.4	216.6	107	0	2	105	16.2	78.1	5.7	216.8
Other	1,014	11	1	1,002	18.0	52.8	29.2	229.1	1,014	12	3	989	11.2	78.6	10.2	221.9
Economic Status																
Economically Disadvantaged	30,946	148	62	30,736	34.3	54.1	11.6	209.8	30,946	123	71	30,415	29.0	68.8	2.2	208.6
Non-Economically Disadvantaged	73,338	158	103	73,077	11.5	54.3	34.2	235.2	73,338	166	98	72,833	7.8	79.9	12.3	224.9

[1] Excludes students who took Braille, large print, alternate or special equated forms.
[2] Percentages may not total 100 due to rounding.
[3] Students are included in Total Students only once, but they appear in all other categories that apply.
[4] Excludes students coded former LEP.
[5] Excludes students who did not have Gender coded.
[6] Students who did not have any Ethnicity coded and students with multiple Ethnicities coded are reported in the Other category only.

Figure 8.2.4: Performance by Demographic Group Statewide

8.4 State Summary Reporting

The statewide student data file summarizes student performance by grade and content areas tested on the total test sections, disaggregated by population, gender, ethnicity, economic status and migrant status. This data file is made available via the NJDOE's Office of Evaluation and Assessment's Web site.

8.5 Interpreting Reports

The NJ ASK score report information is used for the purpose of district monitoring. The data are also provided to assist districts in the review of current curricular programs. With the adoption of the New Jersey Core Curriculum Content Standards in May 1996, all districts were required to implement standards based instruction. NJ ASK results displayed in school-level and district-level reports can provide meaningful information for educational program reviews.

All other factors being equal, the reliability (stability) of scores decreases as the number of items used decreases. Generally speaking, reliability is lower in clusters that have smaller numbers of items. All else being equal, differences in mean cluster scores for clusters with smaller numbers of items must be greater than differences for clusters with large numbers of items before they can be considered meaningful. Decreases in reliability also increase the need for multiple measures, particularly where the number of students in the assessed group is small.

All clusters cannot be assumed to be of equal difficulty level. Cluster scores should, therefore, be compared to their respective Just Proficient Means to facilitate effective interpretation. Insofar as tests are not equated at the cluster level, cluster scores cannot be compared from year to year. Year-to-year comparisons should be limited to total test scores in the content areas tested. For each content area, it is the whole test level (only) for which scores are equated.

The NJ ASK reports provide information on clusters in content areas that need further attention. However, since some clusters were assessed with a relatively small number of items, evaluation of a student's performance should never be based solely on the results of the NJ ASK or any other single form of formal or informal assessment. Insofar as the NJ ASK is equated at the test level only, cluster performance should not be directly compared across multiple test administrations.

PART 9: ACCOUNTABILITY

9.1 Introduction

The 2001 re-authorization of the Elementary and Secondary Education Act of 1965 was signed into federal law January 8, 2002. Characterized in the statute as "An Act to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind," it carries the short title of the No Child Left Behind (NCLB) Act of 2001. Like New Jersey, many states have modified and/or supplemented their student assessments to comply with the federal statute and now use assessment results to make both federal and state accountability decisions. This section of the Technical Report focuses primarily on accountability related to NCLB

9.2 Accountability Model - Overview

Adequate Yearly Progress (AYP) is the term used in NCLB to refer to the minimum improvement required of each school and district over the course of one year. It is measured at the school and district levels by:

- Measuring growth in the percentage of students scoring Proficient or above in reading and mathematics.
- Assessing improvement on one "other academic indicator."
- Testing at least 95% of enrolled students and student subpopulations of sufficient size.

As the term AYP suggests, progress toward NCLB academic goals is evaluated annually. New Jersey's definition of AYP is determined by a formula. The formula calculates the number of proficient scores over the number of valid test scores, with 20% of the items responded to denoting a valid test score. Standards have been set based on starting points and incremental increases aimed at 100 percent proficiency by 2014. Separate starting points for this process have been set for LAL and mathematics for grades 4, 8, and 11.

9.3 Accountability Model - Goals

Using the Compound Annual Growth Rate (CAGR) model, state benchmarks for the standards will be raised every three years in school years 2004–2005, 2007–2008, 2010–2011, and 2013–2014 until the 100 percent proficiency goal is reached for all subject areas at all levels. This methodology employs equal increments of growth calculated on a percentage rate for closing the achievement gap, rather than a straight numerical calculation. Table 9.3.1 presents the annual and intermediate goals based upon application of the CAGR method.

Table 9.3.1
Annual and Intermediate Goals Based Upon Application of CAGR Method

Content Area	Grade	2002-2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2013-2014
LAL	4	68	68	75	75	75	82	82	82	91	91	91	100
	8	58	58	66	66	66	76	76	76	87	87	87	100
	12	73	73	79	79	79	85	85	85	92	92	92	100
Math	4	53	53	62	62	62	73	73	73	85	85	85	100
	8	39	39	49	49	49	62	62	62	79	79	79	100
	12	55	55	64	64	64	74	74	74	86	86	86	100

9.4 Accountability Classification Results

Preliminary AYP status under NCLB accountability requirements for 2006 can be found at the following address:

http://www.nj.gov/njded/title1/accountability/profiles/06/

PART 10: QUALITY CONTROL PROCEDURES

10.1 Quality Control in Data Preparation

Measurement Incorporated (MI) developed a set of rules that data correction staff used to hand-edit information contained on a response document. Data correction provided an opportunity to review and correct the information collected by scanning, ensure that it was scanned accurately, and verify that the information provided by the schools was consistent and correct.

To ensure accuracy, once a correction had been made to the database, the document was validated again to ensure the corrected edit did not create another error. All edits were recorded and tracked in MI's databases, along with the user ID of the staff member making the edits. Performance was constantly and consistently reviewed to ensure that accuracy was maintained and to correct any detected problems.

10.2 Quality Control in Scanning

Scanning and scoring programs were fully tested and reviewed using structured testing methodologies before live test materials were processed, and were continually monitored throughout the process. MI's Quality Assurance (QA) staff developed independent queries to validate all software programs and programmatically produced deliverables for reporting. Each program was tested to ensure that data were included or excluded as appropriate (with particular attention to any special equating situations), and programmatic calculations were performed accurately and according to the reporting rules provided by Riverside and the New Jersey client. During the QA process, reader score sheets were reviewed and compared to student records to verify that scores were applied appropriately. A selection of students was presented to ensure coverage of each type of demographic coding scenario as well as any overrides that were done by MI according to coding rules developed in conjunction with Riverside and the New Jersey client.

MI monitored all aspects of machine scanning. Ensuring the accuracy of demographic data collection was an important component of producing accurate student score reports. Therefore, MI created a detailed data verification plan according to their usual high standards for data capture. This plan encompassed all phases and was a comprehensive set of quality processes to ensure the utmost accuracy of the final reports and file deliverables.

QA staff conducted rigorous tests prior to the scanning of live answer documents to collect student demographic data. Scanning applications that included every scannable document were written using Pearson's ScanTools Plus® application. Each application was tested to ensure it was properly defined and set up. This testing stage was conducted to ensure that the data derived from all grids appearing on the scannable document were included in the export file, were accurately read, and returned the correct value. A quality control sample of answer document demographics (test deck) was created so that all possible responses were verified. This structured method of testing provided exact test

parameters and a methodical way of determining that the output received from the scanner(s) was correct. The documents and the data file created from them were carefully compared to further ensure that results from the scanner were accurate. Accurate scanner calibration was verified at the time of testing, and scanners were re-calibrated to specifications prior to each staff shift change to ensure that calibration remained constant and accurate.

MI has developed a set of comprehensive guidelines for eliminating situations that might threaten the integrity of scanned data. By following these strict guidelines, our scanner operators ensured that the most accurate information possible was read from the document. Scanner operators handled minor response document repairs that allowed the original documents to go through the scanner properly. Small rips in a page were often repaired using cellophane tape, for example. In the rare event that a page from an answer document had more serious damage, the gridded responses from the original, damaged page were transcribed onto a replacement page. A second person verified that the page was transcribed correctly. An adhesive label was placed on the original page explaining that it was transcribed, who transcribed it, and the litho code value (answer folder number) of the page it was transcribed onto. This page was kept with the rest of the document as a reference in case of a question or challenge.

Besides handling student document pages that do not scan, scanner operators also responded to extra pages rejected by the scanner. When an extra page contained a handwritten or typewritten response, the scanner operator filled out a label identifying the document it was associated with and attached that label to the page. The scan bin was set aside, and a scoring assistant was notified. The scoring assistant determined whether the page contained responses that should be used in determining the student's score. If it did, the item with which the extra page was associated was indicated on the label. This extra page was kept with the corresponding original response document page throughout processing so that scoring staff would assign the correct score to the student.

10.3 Quality Control in Editing and Data Input

MI used a successive check of quality assurance and control system to ensure and maintain accurate and timely scoring results, reporting, and dissemination of data. Throughout the execution of the software testing, all defects were logged, assigned, and followed through to resolution. Software changes or "fixes" provided by the developer to resolve defects were re-tested until satisfactory results were achieved. Regression testing of previously tested functionality was performed to ensure that the fix did not adversely affect any other functionality of the application/system.

Deployment of software applications to the staging environment was also tested during the QA process in partnership with MI's Network Operations (NetOps) team. The staging environment closely matched the production environment, which enabled us to determine projected behavior once the application was deployed to the production environment.

10.4 Quality Control in Scoring

MI constantly monitors the quality of each scorer's work throughout every project. Methods used to monitor scorers' scoring habits in scoring NJ ASK included the use of Daily Reader Status Reports.

For writing and open-ended items, each student writing sample was scored holistically by readers using the Registered Holistic Scoring Method. Previously, a different reader from another team read identified packets a second time. At no time did readers see the other's scores. After the scores from each day's work were entered, MI's data application calculated the results and generated a status report. These reports showed the total number of papers read and the percentage agreement of each reader, both perfect and adjacent, for the second-read packets. The reports also showed score point distributions. Scoring directors examined the reports and used the information to determine the need for retraining of individual readers or the group as a whole. It could easily be determined if a reader was consistently scoring "too high" or "too low," as well as the specific score points with which they may have been having difficulty. The Daily Reader Status Reports showed not only the current daily totals for each scorer, but also the project-to-date totals.

Retraining was an ongoing process once scoring began. Daily monitoring of completed packets and analysis of agreement rates provided by the Daily Reader Status Reports and validity packets alerted team leaders and management personnel to individual retraining needs. If it became apparent that a whole team or a whole group was having difficulty with a particular type of response, large group training sessions were conducted. Standard retraining procedures included room-wide discussions led by the scoring director, team discussions conducted by team leaders, spot-checking of individual scorers by team leaders, and discussions between team leaders and individual scorers.

Scorers were dismissed when, in the opinion of the scoring director and the project director, they had been counseled, retrained, and given every reasonable opportunity to improve but were still performing below the acceptable standard.

10.5 Quality Control in Reporting

MI fully recognizes the importance of problem-free score reporting and has employed stringent quality control procedures ensuring that reporting on all levels was complete and accurate to the extent possible for the NJ ASK 5–7 assessment. With this in mind, MI thoroughly tested, reviewed, and proofread all reporting deliverables prior to delivery to Riverside and the New Jersey client.

QA staff verified the content of preliminary reports during the preliminary reporting phase and ensured that reports contained the correct information presented in a clear, concise manner. Reports were tested to ensure that valid values were verified, valid codes were included on student records, correct scores were reflected and were attributed to the correct student, cluster scores were accurately aggregated and totaled, and appropriate student totals were reported in all aggregate reports.

QA also verified formatting of reports, as designed by Riverside and the New Jersey client, including fonts, footnotes, line separations, sections, and headings. This testing process was included in all aspects of data files, electronic reports, and printed reports. During the printing of the final reports, QA verified that print quality was excellent and all reports for all students, schools, and school systems were complete.

PART 11: SUMMARY STATISTICS

11.1 Introduction

This section provides descriptive statistics for number correct raw score and for scale scores. Statistics include N-counts, means, standard deviations, minimum and maximum values, and a variety of data disaggregations, including student demographic group and District Factor Group (DFG).

11.2 Descriptive Statistics for Total Raw Score

Descriptive statistics of total scores of NJ ASK 2006 are summarized in Table 11.2.1 by test content, form, and grade level. A total of 314,316 students participated in the LAL grades 5–7 tests, and 315,775 students participated in the math grades 5–7 tests.

Table 11.2.1
Descriptive Statistics for Total Raw Scores by Content Area and Grade Level

Test	Form	Grade	Ncount	Mean	StdDev	MinObs	MaxObs	Nitem	MaxPoss
LAL	OP	5	103219	22.35	5.97	0	38	25	41
	OP	6	103529	23.47	6.44	0	44	25	48
	OP	7	106748	25.80	6.57	0	46	25	48
	AL	6	376	25.29	6.83	3	40	25	48
	BR	5	9	21.11	5.93	8	26	24	36
	BR	6	8	13.13	6.40	7	25	25	48
	BR	7	4	25.25	7.46	16	34	25	48
	LP	5	93	18.15	6.32	3	31	25	41
	LP	6	78	18.26	8.25	2	36	25	48
	LP	7	72	21.17	7.32	8	37	25	48
Math	OP	5	103809	24.43	7.26	0	39	33	39
	OP	6	104075	21.70	8.33	0	39	33	39
	OP	7	107251	16.47	7.50	0	39	33	39
	AL	6	376	24.34	8.07	0	39	33	39
	BR	5	9	19.33	8.31	5	30	31	37
	BR	6	8	11.63	7.58	4	24	32	38
	BR	7	4	14.25	10.01	4	28	31	35
	LP	5	95	18.34	7.54	2	31	31	37
	LP	6	76	15.78	9.18	0	36	32	38
	LP	7	72	11.72	6.57	3	30	31	35

*OP: Operational Test; AL: Alternative Operational Test; BR: Braille; and LP: Large Print.

11.3 Descriptive Statistics for Total Raw Scores by Cluster

Tables 11.3.1 through 11.3.4 summarize means and standard deviations of raw scores by cluster. Tables 11.3.1 through 11.3.3 refer to the NJ ASK 2006 operational test forms, and Table 11.3.4 refers to the Breach test form. A total of 376 Grade 6 students were administered the Breach test form.

Table 11.3.1
Grade 5 - Means and Standard Deviations for Number Correct Raw Score

	Number of Items		Number of Possible points	Raw Score		Average Percent Correct
	MC	OE		Mean	Standard Deviation	
LAL	**20**	**5**	**41**	**22.35**	**5.97**	**54.50**
Writing		1	5	2.93	0.84	58.63
Reading	20	4	36	19.41	5.51	53.93
Working with Text	12	1	16	9.22	3.13	57.64
Analyzing Text	8	3	20	10.19	2.78	50.96
Math	**30**	**3**	**39**	**24.43**	**7.26**	**62.64**
Number and Numerical Operation	7	1	10	6.46	2.28	64.62
Geometry and Measurement	9	0	9	6.05	1.67	67.19
Patterns and Algebra	7	1	10	5.49	2.23	54.93
Data Analysis, Probability, and Discrete Mathematics	7	1	10	6.43	2.59	64.27
Problem Solving	28	3	37	22.97	6.92	62.09

Table 11.3.2
Grade 6 - Means and Standard Deviations for Number Correct Raw Score

	Number of Items		Number of Possible points	Raw Score		Average Percent Correct
	MC	OE		Mean	Standard Deviation	
LAL	**20**	**5**	**48**	**23.47**	**6.44**	**48.90**
Writing		1	12	5.24	2.06	43.65
Reading	20	4	36	18.23	5.08	50.65
Working with Text	11	1	15	8.47	2.71	56.47
Analyzing Text	9	3	21	9.76	2.84	46.49
Math	**30**	**3**	**39**	**21.70**	**8.33**	**55.65**
Number and Numerical Operation	9	0	9	5.50	2.02	61.09
Geometry and Measurement	7	1	10	5.93	2.49	59.28
Patterns and Algebra	7	1	10	4.61	2.55	46.14
Data Analysis, Probability, and Discrete Mathematics	7	1	10	5.66	2.64	56.62
Problem Solving	28	3	37	20.45	7.98	55.28

Table 11.3.3
Grade 7 - Means and Standard Deviations for Number Correct Raw Score

	Number of Items		Number of Possible points	Raw Score		Average Percent Correct
	MC	OE		Mean	Standard Deviation	
LAL	**20**	**5**	**48**	**25.80**	**6.57**	**53.74**
Writing		1	12	5.54	1.97	46.17
Reading	20	4	36	20.26	5.29	56.27
Working with Text	12	0	12	8.55	2.37	71.22
Analyzing Text	8	4	24	11.71	3.39	48.79
Math	**30**	**3**	**39**	**16.47**	**7.50**	**42.24**
Number and Numerical Operation	7	1	10	4.41	2.51	44.08
Geometry and Measurement	9	0	9	4.29	1.93	47.62
Patterns and Algebra	7	1	10	3.50	2.39	35.00
Data Analysis, Probability, and Discrete Mathematics	7	1	10	4.28	2.07	42.79
Problem Solving	28	3	37	15.28	7.16	41.29

Table 11.3.4
Grade 6 – Breach Form
Means and Standard Deviations for Number Correct Raw Score

	Number of Items		Number of Possible points	Raw Score		Average Percent Correct
	MC	OE		Mean	Standard Deviation	
LAL	**20**	**5**	**48**	**25.29**	**6.83**	**52.69**
Writing		1	12	5.78	2.05	48.18
Reading	20	4	36	19.51	5.30	54.20
Working with Text	11	1	15	8.64	2.75	57.62
Analyzing Text	9	3	21	10.87	3.02	51.75
Math	**30**	**3**	**39**	**24.34**	**8.07**	**62.40**
Number and Numerical Operation	9	0	9	6.10	1.86	67.76
Geometry and Measurement	7	1	10	6.75	2.29	67.47
Patterns and Algebra	7	1	10	5.17	2.48	51.70
Data Analysis, Probability, and Discrete Mathematics	7	1	10	6.32	2.59	63.19
Problem Solving	28	3	37	22.97	7.78	62.09

11.4 Scale Score Distributions by Content Area and Grade

Descriptive statistics of scale scores and percentage distributions of students' performance levels are summarized in Table 11.4.1 by content area and grade. For LAL, students that were flagged as "LAL void" or "LAL not present" (LAL Scale Score marked as "0") were removed. For mathematics, students that were flagged as "Math void" or "math not present" (Math Scale Score marked as "0") were removed. For all test forms, scale scores have a range of 100 to 300. A student is classified as Partially Proficient (PP) if his/her scale score is lower than 200. A student is classified as Advanced Proficient (AP) if his/her scale score is 250 or higher. The other students are classified as Proficient (P).

Table 11.4.1
Descriptive Statistics of Scale Scores and Percentage Distributions of Students' Performance Levels by Content area and Grade

Test	Grade	Form	N	Mean	StdDev	Min	Max	%PP	%P	%AP
LAL	5	Operational	103219	220.10	20.49	110	300	14.05	76.63	9.33
	6	Operational	103529	213.32	27.05	100	300	25.04	65.80	9.16
	7	Operational	106748	217.10	24.41	100	300	19.75	70.35	9.90
	6	Alternative	376	221.24	29.95	108	300	17.29	65.96	16.76
	5	Braille	9	225.67	22.92	179	250	11.11	66.67	22.22
	6	Braille	8	170.50	27.18	142	217	87.50	12.50	0.00
	7	Braille	4	215.75	26.51	186	250	25.00	50.00	25.00
	5	Large Print	93	206.26	18.90	154	257	33.33	65.59	1.08
	6	Large Print	78	192.01	36.06	100	279	60.26	32.05	7.69
	7	Large Print	72	201.25	25.19	155	272	43.06	52.78	4.17
Math	5	Operational	103809	227.66	34.45	100	300	18.24	54.24	27.52
	6	Operational	104075	216.90	32.48	100	300	29.26	53.49	17.26
	7	Operational	107251	212.29	32.36	100	300	35.71	49.85	14.44
	6	Alternative	376	227.18	33.26	100	300	16.76	58.51	24.73
	5	Braille	9	208.44	40.96	129	259	33.33	55.56	11.11
	6	Braille	8	178.25	31.13	142	226	75.00	25.00	0.00
	7	Braille	4	206.75	51.81	149	275	25.00	50.00	25.00
	5	Large Print	95	204.19	37.80	100	265	35.79	52.63	11.58
	6	Large Print	76	195.47	37.70	100	290	56.58	32.89	10.53
	7	Large Print	72	195.33	32.92	138	289	63.89	23.61	12.50

11.5 Scale Score Distributions by Demographic Group

Descriptive statistics of scale scores and percentage distributions of students' performance levels by demographic groups are summarized in Tables 11.5.1 through 11.5.6 by content area and grade. For LAL, students that were flagged as "LAL void" or "LAL not present" (LAL Scale Score marked as "0") were removed. For mathematics, students that were flagged as "Math void" or "math not present" (Math Scale Score marked as "0") were removed. Students flagged as "Special Education" (SpeEdu) or "Limited English Proficient" (LEP) were removed from the "General Education" (GenEdu) group.

Table 11.5.1
LAL Grade 5 - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by Demographic Groups

DemoGrp	N	Mean	STD	Min	Max	%PP	%P	%AP
GenEdu	84704	224.05	18.73	110	300	7.58	81.31	11.10
SpeEdu	16090	202.75	18.48	110	300	41.79	56.94	1.27
LEP	2602	196.82	17.16	110	265	57.07	42.24	0.69
Female	50168	221.94	20.50	110	300	11.69	77.23	11.08
Male	53039	218.36	20.33	110	300	16.28	76.05	7.67
Asian	7538	230.38	20.63	110	300	5.60	73.53	20.87
Black	18040	208.10	18.49	110	300	30.39	67.36	2.26
Hispanic	18287	210.97	18.84	110	300	24.87	72.03	3.10
Native	105	216.83	20.03	161	300	16.19	78.10	5.71
Pacific	309	228.25	20.40	170	300	6.15	77.02	16.83
White	57951	225.31	18.74	110	300	6.73	81.33	11.95
Other	989	221.91	21.03	132	300	11.22	78.56	10.21
NotEcnDis	72814	224.88	19.41	110	300	7.80	79.89	12.31
EcnDis	30405	208.65	18.37	110	300	29.01	68.80	2.19
NonMigrant	103191	220.10	20.49	110	300	14.04	76.63	9.33
Migrant	28	206.04	19.09	170	257	35.71	60.71	3.57
NonTitle1	83018	223.39	19.90	110	300	9.87	78.90	11.23
Title 1	20201	206.60	17.07	110	300	31.21	67.28	1.50
Sect 504A	14968	202.63	17.99	110	300	41.49	57.48	1.03
Sect 504B	14870	202.59	18.03	110	300	41.53	57.47	1.00
Sect 504C	749	202.30	18.41	154	275	43.66	54.74	1.60
Sect 504D	13983	202.34	17.85	110	287	41.97	57.13	0.91

Table 11.5.2
LAL Grade 6 - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by Demographic Groups

DemoGrp	N	Mean	STD	Min	Max	%PP	%P	%AP
GenEdu	85356	218.85	23.86	100	300	16.79	72.26	10.96
SpeEdu	16095	188.31	25.86	100	298	62.47	36.78	0.75
LEP	2228	179.56	26.14	100	298	75.54	24.10	0.36
Female	50361	216.26	26.67	100	300	21.87	67.02	11.11
Male	53010	210.58	27.08	100	300	28.02	64.67	7.31
Asian	7240	227.75	26.97	100	300	10.80	67.56	21.64
Black	18469	197.09	25.49	100	300	48.42	49.76	1.82
Hispanic	17861	201.00	25.45	100	300	40.99	56.43	2.58
Native	120	212.68	24.93	153	298	28.33	61.67	10.00
Pacific	210	219.80	23.23	135	288	13.33	75.71	10.95
White	58532	220.46	23.96	100	300	14.50	73.58	11.92
Other	1097	210.20	30.51	100	300	30.45	60.26	9.30
NotEcnDis	73572	219.61	25.14	100	300	16.28	71.61	12.11
EcnDis	29957	197.89	25.31	100	300	46.55	51.54	1.90
NonMigrant	103492	213.33	27.04	100	300	25.03	65.81	9.16
Migrant	37	194.22	26.05	135	250	62.16	35.14	2.70
NonTitle1	83487	217.42	26.10	100	300	19.22	69.72	11.06
Title 1	20042	196.25	24.04	100	300	49.29	49.49	1.22
Sect 504A	15233	188.90	25.52	100	288	61.67	37.56	0.77
Sect 504B	15078	189.01	25.58	100	288	61.49	37.70	0.82
Sect 504C	629	193.88	24.79	108	288	54.69	44.04	1.27
Sect 504D	13926	188.34	25.28	100	288	62.78	36.55	0.67

Table 11.5.3
LAL Grade 7 - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by Demographic Groups

DemoGrp	N	Mean	STD	Min	Max	%PP	%P	%AP
GenEdu	87814	222.18	21.66	100	300	11.77	76.33	11.89
SpeEdu	16684	194.46	22.29	100	291	55.09	44.20	0.70
LEP	2382	186.20	23.29	100	257	69.69	30.10	0.21
Female	51823	221.51	24.16	100	300	14.93	71.71	13.36
Male	54788	212.95	23.90	100	300	24.27	69.10	6.63
Asian	7203	229.87	24.46	108	300	7.97	70.19	21.84
Black	19019	202.31	22.38	100	300	41.21	56.70	2.09
Hispanic	18287	205.77	22.94	100	300	33.67	63.42	2.91
Native	142	210.71	23.74	130	272	24.65	69.01	6.34
Pacific	240	221.00	20.86	159	300	12.92	76.67	10.42
White	60901	223.65	21.84	100	300	10.22	76.74	13.05
Other	956	214.80	25.38	100	300	23.85	67.36	8.79
NotEcnDis	77324	222.54	22.77	100	300	12.25	74.92	12.83
EcnDis	29424	202.82	22.74	100	300	39.46	58.33	2.20
NonMigrant	106722	217.11	24.41	100	300	19.74	70.36	9.90
Migrant	26	195.15	23.34	150	257	50.00	46.15	3.85
NonTitle1	86893	220.70	23.57	100	300	14.70	73.51	11.80
Title 1	19855	201.34	21.60	100	300	41.86	56.54	1.60
Sect 504A	15692	195.36	22.45	100	300	53.49	45.64	0.87
Sect 504B	15745	195.27	22.53	100	300	53.60	45.53	0.88
Sect 504C	688	197.10	24.18	100	291	50.29	47.38	2.33
Sect 504D	13885	194.43	22.16	100	300	55.20	44.11	0.68

Table 11.5.4
Mathematics Grade 5 - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by Demographic Groups

DemoGrp	N	Mean	STD	Min	Max	%PP	%P	%AP
GenEdu	84662	234.03	31.02	100	300	11.64	56.35	32.01
SpeEdu	16068	200.32	34.85	100	300	46.12	45.90	7.98
LEP	3264	194.65	33.69	100	300	54.66	39.43	5.91
Female	50456	227.84	33.01	100	300	17.09	56.09	26.81
Male	53340	227.48	35.76	100	300	19.31	52.49	28.20
Asian	7639	248.43	32.74	100	300	6.68	41.71	51.62
Black	18013	206.72	32.96	100	300	37.91	52.29	9.80
Hispanic	18722	213.89	32.98	100	300	29.53	56.11	14.36
Native	105	216.59	31.94	114	300	24.76	63.81	11.43
Pacific	308	242.79	32.43	114	300	7.14	48.70	44.16
White	58020	235.78	30.72	100	300	10.06	55.93	34.01
Other	1002	229.10	35.76	100	300	17.96	52.79	29.24
NotEcnDis	73075	235.16	32.23	100	300	11.47	54.32	34.21
EcnDis	30734	209.82	32.97	100	300	34.32	54.05	11.63
NonMigrant	103775	227.66	34.45	100	300	18.23	54.24	27.53
Migrant	34	208.21	36.63	114	295	35.29	55.88	8.82
NonTitle1	83408	232.66	33.18	100	300	13.67	54.35	31.98
Title 1	20401	207.19	31.88	100	300	36.90	53.79	9.31
Sect 504A	14959	200.25	34.02	100	300	45.77	46.96	7.27
Sect 504B	14859	200.21	34.03	100	300	45.72	47.00	7.28
Sect 504C	748	197.72	35.13	100	300	48.80	44.12	7.09
Sect 504D	13971	199.86	34.04	100	300	46.23	46.61	7.16

Table 11.5.5
Mathematics Grade 6 - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by Demographic Groups

DemoGrp	N	Mean	STD	Min	Max	%PP	%P	%AP
GenEdu	85309	222.96	30.33	100	300	21.05	58.52	20.43
SpeEdu	16050	189.34	27.18	100	300	66.61	30.63	2.76
LEP	2875	188.80	27.29	100	300	67.34	29.29	3.37
Female	50631	216.97	31.72	100	300	28.62	54.59	16.79
Male	53284	216.86	33.17	100	300	29.83	52.46	17.71
Asian	7362	239.07	32.80	100	300	10.98	48.59	40.44
Black	18404	196.38	27.60	100	300	55.30	40.57	4.13
Hispanic	18226	204.01	28.68	100	300	43.67	49.36	6.97
Native	121	212.80	33.34	118	300	37.19	48.76	14.05
Pacific	217	227.76	31.05	154	300	17.97	53.92	28.11
White	58627	224.61	30.13	100	300	18.82	59.55	21.63
Other	1118	212.87	35.07	100	300	34.62	48.21	17.17
NotEcnDis	73853	223.88	31.48	100	300	20.77	57.11	22.12
EcnDis	30222	199.86	28.32	100	300	50.00	44.63	5.37
NonMigrant	104033	216.91	32.48	100	300	29.24	53.49	17.26
Migrant	42	193.19	27.29	148	265	61.90	33.33	4.76
NonTitle1	83838	221.59	31.98	100	300	23.37	56.18	20.45
Title 1	20237	197.47	26.84	100	300	53.64	42.33	4.03
Sect 504A	15190	189.80	26.98	100	300	65.98	31.26	2.76
Sect 504B	15034	189.89	27.09	100	300	65.78	31.38	2.83
Sect 504C	628	194.42	28.45	100	291	57.32	38.85	3.82
Sect 504D	13889	189.13	26.61	100	300	66.98	30.49	2.53

Table 11.5.6
Mathematics Grade 7 - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by Demographic Groups

DemoGrp	N	Mean	STD	Min	Max	%PP	%P	%AP
GenEdu	87692	218.06	30.96	100	300	27.50	55.27	17.23
SpeEdu	16596	185.95	24.72	100	300	73.10	25.17	1.73
LEP	3103	188.64	26.54	100	300	69.93	27.07	3.00
Female	52082	211.30	31.11	100	300	36.11	51.07	12.83
Male	55033	213.24	33.48	100	300	35.31	48.72	15.97
Asian	7325	237.40	34.30	122	300	13.41	48.10	38.50
Black	18918	192.39	25.57	100	300	62.74	34.58	2.69
Hispanic	18730	199.46	27.17	100	300	50.96	44.17	4.87
Native	141	210.23	31.04	122	300	36.88	49.65	13.48
Pacific	245	222.91	30.02	122	300	18.37	60.41	21.22
White	60930	219.39	30.80	100	300	25.31	56.58	18.11
Other	962	210.35	32.86	122	300	39.81	45.32	14.86
NotEcnDis	77579	218.75	32.03	100	300	27.28	54.16	18.56
EcnDis	29672	195.40	26.63	100	300	57.73	38.60	3.67
NonMigrant	107223	212.30	32.36	100	300	35.70	49.86	14.44
Migrant	28	188.89	25.23	136	258	67.86	28.57	3.57
NonTitle1	87180	216.63	32.30	100	300	29.84	53.02	17.14
Title 1	20071	193.43	25.05	100	300	61.20	36.10	2.70
Sect 504A	15612	186.97	25.25	100	300	71.73	26.24	2.04
Sect 504B	15661	186.90	25.24	100	300	71.73	26.22	2.05
Sect 504C	683	188.60	24.84	100	285	69.11	29.14	1.76
Sect 504D	13819	185.88	24.70	100	300	73.38	24.97	1.66

11.6 Scale Score Distributions by District Factor Groups (DFG)

New Jersey has a long history of using District Factor Groups (DFG)[23] in the analysis of assessment results. The statistical method for developing DFGs includes U.S. Census data and has been improved over the years to provide a better measure of social economic status (SES). The DFGs included in the tables below are A, B, CD, DE, FG, GH, I, J, R, and V. Note that group "R" in the District Factor Groups is charter schools and group "V" is vocational schools.

Descriptive statistics for scale scores and percentage distributions of students' performance levels by DFG are summarized in Tables 11.6.1 and 11.6.2 by content area and grade. For LAL, students that were flagged as "LAL void" or "LAL not present" (LAL Scale Score marked as "0") were removed. For mathematics, students that were flagged as "Math void" or "math not present" (Math Scale Score marked as "0") were removed.

[23] For more information on DFGs, see the following link: http://www.state.nj.us/njded/finance/sf/dfg.shtml.

Table 11.6.1
LAL - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by DFG

Grade	DFG	N	Mean	StdDev	Min	Max	%PP	%P	%AP
5	A	17796	206.49	18.57	110	300	33.59	64.42	1.98
	B	10686	214.46	18.80	110	300	18.95	76.53	4.52
	CD	8823	216.54	18.75	132	300	16.28	78.41	5.32
	DE	15610	221.23	18.82	110	300	10.20	81.29	8.50
	FG	12848	223.14	18.75	132	300	8.31	81.93	9.75
	GH	14277	225.21	19.06	110	300	7.44	80.25	12.31
	I	19834	230.10	19.05	154	300	4.18	77.99	17.83
	J	1798	232.66	18.93	154	300	2.45	75.31	22.25
	R	1492	209.97	18.34	146	300	28.35	68.36	3.28
6	A	17729	194.28	26.47	100	300	52.35	46.02	1.63
	B	10825	206.1	24.47	100	300	33.83	62.05	4.12
	CD	9090	208.88	24.47	100	300	28.9	65.91	5.19
	DE	15515	213.97	23.78	100	300	21.39	71.57	7.04
	FG	12735	217.64	24.46	100	300	18.15	71.52	10.33
	GH	14274	221.5	24.7	100	300	14.56	72	13.44
	I	20005	226.75	24.19	100	300	9.19	72.95	17.86
	J	1735	228.3	23.22	108	300	7.03	74.29	18.67
	R	1551	201.45	24.24	108	300	41.72	55.58	2.71
	V	23	236.3	21.41	193	300	4.35	69.57	26.09
7	A	17852	199.74	23.07	100	300	45.33	53.09	1.57
	B	11101	210.04	22.56	100	300	27.30	68.13	4.57
	CD	10059	212.89	21.97	100	300	22.28	72.30	5.42
	DE	16044	217.99	21.87	100	300	16.24	74.93	8.83
	FG	13358	220.45	21.92	100	300	13.34	76.37	10.29
	GH	14296	225.10	22.53	100	300	9.70	75.41	14.89
	I	20799	229.04	21.86	100	300	6.38	75.05	18.58
	J	1730	232.57	21.48	137	300	3.58	72.95	23.47
	R	1423	206.31	21.18	121	300	35.21	61.98	2.81
	V	14	231.36	18.21	209	281	85.71	14.29	0.00

Table 11.6.2
Mathematics - Descriptive Statistics for Scale Scores and Percentage Distributions of Students' Performance Levels by DFG

Grade	DFG	N	Mean	StdDev	Min	Max	%PP	%P	%AP
5	A	18070	206.06	33.93	100	300	38.96	50.78	10.26
	B	10776	220.40	32.76	100	300	22.98	57.53	19.50
	CD	8868	222.46	32.67	100	300	20.82	58.19	21.00
	DE	15656	230.39	31.16	100	300	13.82	58.36	27.82
	FG	12896	232.04	31.15	100	300	12.60	57.37	30.03
	GH	14310	235.95	31.80	100	300	10.92	54.46	34.63
	I	19883	242.50	31.26	100	300	7.43	49.63	42.94
	J	1804	245.98	30.67	133	300	5.65	48.28	46.06
	R	1490	208.31	34.38	114	300	40.13	46.38	13.49
6	A	17955	196.41	28.44	100	300	55.67	39.63	4.7
	B	10902	209.7	29.59	100	300	36.09	53.51	10.4
	CD	9112	213.14	30.24	100	300	31.66	55.81	12.53
	DE	15553	215.89	29.33	100	300	27.53	58.75	13.72
	FG	12779	220.92	30.98	100	300	23.48	57.28	19.23
	GH	14336	225.52	31.24	100	300	19.04	57.63	23.33
	I	20080	232.66	30.9	100	300	13	56	31
	J	1740	235.27	29.7	118	300	10.4	56.72	32.87
	R	1548	200.56	30.39	131	300	50.84	42.31	6.85
	V	23	230.26	21.16	165	280	4.35	86.96	8.7
7	A	18140	192.89	26.38	100	300	62.08	34.77	3.14
	B	11156	203.89	28.09	100	300	44.85	48.29	6.86
	CD	10062	206.76	28.77	100	300	40.39	50.95	8.66
	DE	16034	211.29	29.74	100	300	34.75	53.38	11.87
	FG	13373	214.54	30.82	100	300	30.99	54.65	14.36
	GH	14420	221.69	32.01	100	300	23.70	55.23	21.07
	I	20832	228.65	32.48	100	300	17.83	54.36	27.81
	J	1731	232.58	31.71	122	300	14.33	54.48	31.20
	R	1419	196.96	28.62	122	300	56.94	37.21	5.85
	V	14	218.64	35.17	147	285	21.43	64.29	14.29

PART 12: RELIABILITY

12.1 Introduction

The New Jersey Department of Education is required by federal law to ensure that the instruments it uses to measure student achievement for school accountability provide reliable results. This section shows that results of the 2006 NJ ASK for grades 5–7 measure student achievement in a reliable manner. The size of the measurement error associated with test scores is reasonable and can be taken into account when interpreting the scores of individual students.

12.2 Reliability and Measurement Error

Reliable student test scores, like other reliable measurements, are consistent. More specifically, measurement components are consistent with each other. Results of the components vary, but they do so within tolerable limits. In general, measurement error and reliability are inversely related. When measurement error is large, reliability is small. Increasing reliability by minimizing measurement error is an important goal in the construction of any test.

The NJ ASK assessments, like many other standardized achievement tests, were designed under the assumptions of Classical Test Theory (CTT). This approach builds on the notion of an ideal, error-free or true measurement score. Any observed measurement, such as test score X, is defined as a composite of true score T and its associated error:

$$X = T + \text{error}$$

Estimating the size of the measurement error associated with the true score is the key to estimating reliability. Errors in measurement can result from any of a multitude of factors, including environmental factors (e.g., testing conditions) and examinee factors (e.g., fatigue, stress). Feldt and Brennan (1989)[24] note, "Quantification of the consistency and inconsistency in examinee performance constitutes the essence of reliability analysis" (p. 105). CTT provides a means for this quantification of examinee inconsistency (i.e., measurement error).

The definitions or assumptions in CTT lead to several important properties. For example, it can be demonstrated that

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2,$$

or observed score variance equals the sum of true score variance plus error variance. The relationships among variance terms (i.e., $\sigma_x^2, \sigma_t^2, \sigma_e^2$) are critical to a more thorough understanding of important CTT concepts, including reliability and the standard error of

[24] Feldt and Brennan (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.). Washington, DC: American Council on Education.

measurement. For example, CTT reliability is defined as the correlation between observed scores on parallel forms, which is equal to

$$\rho_{x_1 x_2} = \sigma_t^2 / \sigma_x^2.$$

Reliability in CTT is thus conceptualized as true score variance divided by observed score variance. With just a few algebraic steps, the CTT definition of the standard error of measurement (SEM) can be shown as

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{x_1 x_2}}.$$

Although the conceptualization of reliability and SEM is relatively straightforward, issues underlying the estimation of reliability are not. Reliability can be estimated via the correlation of scores on parallel forms or from test-retest data, or it can be estimated from a single test administration using any one of a variety of techniques (e.g., Brown, 1910; Cronbach, 1951; Kuder & Richardson, 1937)[25]. A very popular technique for estimating reliability from a single test administration is Cronbach's coefficient alpha.

12.3 Test Metrics and Units of Analysis

The NJ ASK quantifies student achievement on three different metrics: number correct raw score, IRT scale, and performance score. While it is the knowledge and skills of individual students that are measured, student scores are aggregated and disaggregated into various units (e.g., school by grade, student group by grade, school, district, and state). Measurement error specific to each metric and each unit of analysis is taken into account when results are reported and accountability decisions are made. It is the responsibility of test developers to maximize reliability and minimize error by (1) identifying likely sources of error; (2) controlling the conditions of error; (3) estimating the size of error and/or level of reliability; and (4) reporting the estimates by metric and unit of analysis.

12.4 Sources of Measurement Error

The scoring of student responses to multiple-choice items is done electronically. Scoring error may result from improper coding and extraneous marks on scannable response sheets. The size of this sort of error is usually small and is controlled though proper test administration procedures, including instructions on how to fill out response sheets and how to erase extraneous markings. The test contractor, who manages the scoring, also uses procedures to minimize this error.

[25] Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*, 296–322. Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334. Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151–160.

Measurement Incorporated (MI) performs a *cusp check,* or multiple choice verification process, for any student whose bubbling errors, if corrected, would give them a passing score. This involves identifying the affected multiple-choice (MC) answer pages and physically reviewing each one for each student in this group. Two of the most common types of errors are use of pen rather than pencil and double-bubbling, which often turns out to be just a bad erasure. If the student's responses are clear to the human eye, as opposed to the machine's eye, the correct response is recorded and the score is changed.

Open-ended items are susceptible to scoring error due to ambiguity in scoring rubrics as well as to differences among raters. Rubrics must be written to balance generality and specificity, covering the range of student responses, while at the same time allowing raters to easily identify the response characteristics distinguishing each score category. To minimize error due to raters, the test contractor thoroughly trains raters and monitors the scoring process. Only raters who meet the contractor's criteria for consistent scoring during training are retained as scorers. The contractor monitors scoring by routinely computing and recording inter-rater agreement.

12.5 Evidence of Raw Score Internal Consistency

Consistency of individual student performance was estimated using coefficient alpha. As previously noted, coefficient alpha is conceptualized as the proportion of total raw score variance that may be attributed to a student's true score variance. Ideally, more score variance should be attributable to true test scores than to measurement error. Alpha is an appropriate index of internal consistency for use on untimed tests such as NJ ASK.

Separate analyses were performed for each grade level and content area. Both multiple-choice and open-ended items scores were used in the computations. Coefficient alpha can be interpreted as a lower bound to reliability and was estimated using the following formula:

$$\alpha_{\text{Cronbach}} = \frac{n}{n-1}[1 - \frac{\sum_{i=1}^{n} \sigma_{Y_i}^2}{\sigma_X^2}],$$

where n is the number of items, $\sigma_{Y_i}^2$ is the variance of item i, and σ_X^2 is the variance of total score. SEM can be interpreted as "the square root of the average of the person-specific error variances of all examinees who participated in the reliability estimation experiment" (Traub, 1994, p. 114[26]). SEMs were calculated using the following formula:

$$SEM = S_X \sqrt{1 - \alpha_{\text{Cronbach}}},$$

where S_X is the standard deviation of observed total scores. Table 12.5.1 summarizes coefficient alpha and SEMs by content and form. Tables 12.5.2 through 12.5.5 summarize coefficient alpha and SEMs of content clusters by test.

[26] Traub, R. E. (1994). Reliability for the social sciences, v3. Thousand Oaks, Sage.

Table 12.5.1
Summary of Coefficient Alpha and SEM by Grade and Content Area

Test	Form*	Grade	Ncount	Cronbach Alpha	SEM
LAL	OP	5	103219	0.85	2.28
	OP	6	103529	0.78	2.99
	OP	7	106748	0.80	2.91
	AL	6	376	0.82	2.93
Math	OP	5	103809	0.84	2.91
	OP	6	104075	0.88	2.87
	OP	7	107251	0.86	2.83
	AL	6	376	0.89	2.73

*OP: Operational Test; AL, Alternative Test. N-counts were insufficient to produce values for Braille and Large Print.

Table 12.5.2
Grade 5 - Coefficient Alpha and Standard Error Measurement for Clusters

	Number of Items		Number of Possible points	Alpha	SEM
	MC	OE			
LAL	**20**	**5**	**41**	**0.85**	**2.28**
Writing		1	5		
Reading	20	4	36	0.85	2.15
Working with Text	12	1	16	0.74	1.59
Analyzing Text	8	3	20	0.73	1.46
Math	**30**	**3**	**39**	**0.84**	**2.91**
Number and Numerical Operation	7	1	10	0.58	1.48
Geometry and Measurement	9	0	9	0.46	1.23
Patterns and Algebra	7	1	10	0.56	1.47
Data Analysis, Prob., & Discrete Math.	7	1	10	0.62	1.60
Problem Solving	28	3	37	0.83	2.86

Table 12.5.3
Grade 6 - Coefficient Alpha and Standard Error Measurement for Clusters

	Number of Items		Number of Possible points	Alpha	SEM
	MC	OE			
LAL	**20**	**5**	**48**	**0.78**	**2.99**
Writing		1	12		
Reading	20	4	36	0.80	2.27
Working with Text	11	1	15	0.68	1.54
Analyzing Text	9	3	21	0.65	1.68
Math	**30**	**3**	**39**	**0.88**	**2.87**
Number and Numerical Operation	9	0	9	0.61	1.26
Geometry and Measurement	7	1	10	0.68	1.40
Patterns and Algebra	7	1	10	0.65	1.51
Data Analysis, Prob., & Discrete Math.	7	1	10	0.66	1.53
Problem Solving	28	3	37	0.88	2.81

Table 12.5.4
Grade 7 - Coefficient Alpha and Standard Error Measurement for Clusters

	Number of Items		Number of Possible points	Alpha	SEM
	MC	OE			
LAL	**20**	**5**	**48**	**0.80**	**2.91**
Writing		1	12		
Reading	20	4	36	0.82	2.24
Working with Text	12	0	12	0.64	1.42
Analyzing Text	8	4	24	0.74	1.72
Math	**30**	**3**	**39**	**0.86**	**2.83**
Number and Numerical Operation	7	1	10	0.61	1.56
Geometry and Measurement	9	0	9	0.54	1.31
Patterns and Algebra	7	1	10	0.63	1.45
Data Analysis, Prob., & Discrete Math.	7	1	10	0.59	1.32
Problem Solving	28	3	37	0.85	2.76

Table 12.5.5
Grade 6 Breach - Coefficient Alpha and Standard Error Measurement for Clusters

	Number of Items		Number of Possible points	Alpha	SEM
	MC	OE			
LAL	**20**	**5**	**48**	**0.82**	**2.93**
Writing		1	12		
Reading	20	4	36	0.82	2.22
Working with Text	11	1	15	0.71	1.64
Analyzing Text	9	3	21	0.71	1.49
Math	**30**	**3**	**39**	**0.89**	**2.73**
Number and Numerical Operation	9	0	9	0.59	1.19
Geometry and Measurement	7	1	10	0.67	1.31
Patterns and Algebra	7	1	10	0.63	1.50
Data Analysis, Prob., & Discrete Math.	7	1	10	0.69	1.44
Problem Solving	28	3	37	0.88	2.69

12.6 Evidence Supporting Rater Reliability

Tables 12.6.1, 12.6.2, and 12.6.3 show the percentages of writing tasks and open-ended items scored with exact agreement, adjacent agreement, and resolution needed by grade level and content area. The score rubrics used for raters had a score range of 0 to 5 for the grade 5 writing prompt, and 0 to 6 for the grade 6 and 7 writing prompt. For grade 6 and 7, the writing scores were doubled in data analyses and score reporting. The rubrics had score points that ranged from 0 to 4 for the LAL open-ended items and from 0 to 3 for the mathematics open-ended items. There were no half points assigned for the writing and open-ended items.

Ten percent (10%) of the writing and open-ended responses were read by a second rater. The purpose of a second-reading for the writing and open-ended responses is to investigate the consistency between raters in the NJ ASK 2006 project. For grade 5 LAL, over 76% of the responses were assigned a score by a second rater that was in exact agreement with the first rater. Another 24% of the second ratings were assigned an adjacent score by a second rater. An adjacent score is a score assigned by the second rater that is no more than ±1 score point from the score assigned by the first rater. For grade 5 mathematics, over 88% of the responses were assigned a score by a second rater that was in exact agreement with the first rater. The results for grades 6 and 7 were comparable, with the exception of grade 7 LAL where the percentage of exact agreement was lower than the other three grades.

Table 12.6.1
Grade 5 - Consistency between Rater Scoring for the Writing Task and Open-Ended Items

	% Raters in Exact Agreement	% Raters in Adjacent Agreement	% Resolution Needed
LAL All	75.50	24.00	0.40
Writing	70.60	28.40	0.80
Reading All	76.80	22.90	0.40
Open-ended 1	75.60	24.20	0.20
Open-ended 2	75.30	24.40	0.20
Open-ended 3	78.70	21.20	0.20
Open-ended 4	77.40	21.80	0.80
Math All	88.40	9.80	1.50
Open-ended 1	84.50	14.20	1.00
Open-ended 2	86.70	10.80	2.40
Open-ended 3	94.10	4.40	1.20

Table 12.6.2
Grade 6 - Consistency between Rater Scoring for the Writing Task and Open-Ended Items

	% Raters in Exact Agreement	% Raters in Adjacent Agreement	% Resolution Needed
LAL All	69.20	29.60	1.10
Writing	60.50	35.60	3.80
Reading All	71.40	28.20	0.50
Open-ended 1	71.80	27.60	0.60
Open-ended 2	70.30	29.00	0.60
Open-ended 3	72.30	27.40	0.40
Open-ended 4	71.30	28.60	0.20
Math All	86.00	12.50	1.20
Open-ended 1	86.80	11.20	2.00
Open-ended 2	91.20	8.20	0.20
Open-ended 3	80.10	18.00	1.40

Table 12.6.3
Grade 7 - Consistency between Rater Scoring for the Writing Task and Open-Ended Items

	% Raters in Exact Agreement	% Raters in Adjacent Agreement	% Resolution Needed
LAL All	62.00	33.90	4.10
Writing	59.80	36.80	3.20
Reading All	62.60	33.00	4.40
Open-ended 1	66.90	31.80	1.30
Open-ended 2	65.50	32.70	1.20
Open-ended 3	58.80	37.10	3.60
Open-ended 4	59.10	31.30	11.40
Math All	91.80	7.30	0.60
Open-ended 1	89.00	9.80	0.80
Open-ended 2	91.80	7.60	0.40
Open-ended 3	94.50	4.60	0.60

12.7 Conditional Estimate of Error at Each Cut-Score

The NJ ASK 2006 grades 5, 6, and 7 cut scores and the corresponding conditional standard error of measurement (CSEM) approved by the New Jersey State Board of Education on May 17, 2006, are summarized in Table 12.7.1. The CSEM is calculated as:

$CSEM = (SD_{obs} / SD_{adj'd}) * SE_{theta}$

where: SD_{obs} is the observed standard deviation for number correct raw score,
$SD_{adj'd}$ is an estimate of the "true" sample standard deviation, and
SE_{theta} is the standard error for theta.

Table 12.7.1
NJ ASK 2006 Cut Scores with Conditional Standard Error of Measurement by Content Area and Grade Level

		Language Arts Literacy		Mathematics	
		Proficient	Advanced Proficient	Proficient	Advanced Proficient
Grade 5	Cut score*	16	30	18	30
	CSEM	1.7	2.7	2.5	2.7
Grade 6	Cut score*	20	32	17	31
	CSEM	1.5	2.0	3.8	4.4
Grade 7	Cut score*	21	34	13	26
	CSEM	1.5	2.2	3.4	3.2

*Cut scores were approved by the New Jersey State Board of Education on May 17, 2006.

12.8 Reliability of Classifications

Reliability indices for proficiency classifications (kappa) were computed with the BB-CLASS program (Brennan, 2004)[27] that is based on the beta-binomial model. Coefficient kappa is given by

$$\kappa = \frac{\varphi - \varphi_c}{1 - \varphi_c},$$

where φ is the probability of a consistent classification and φ_c is the probability of a consistent classification by chance. A classification consistency index can be regarded as the percentage of examinees that would be assigned to, hypothetically, the same achievement level if the same test was administered a second time or an equivalent test

[27] Brennan, R. L. (2004). Manual for BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy (version 1). CASMA Research Report 9. Iowa City, IA.

was administered under the same conditions. Coefficient kappa depends on test score variability, test length, and achievement levels as well as the position of cut-scores (Huynh, 1976)[28].

Coefficients kappa and consistency classification indices are summarized in Table 12.8.1. Kappa and the consistency indexes were estimated for three achievement levels (PP, P, and AP) and two achievement levels (Pass or Fail). A student is regarded as "Pass" if one's achievement level is "P" or "AP" and as "Fail" if one's achievement level is "PP". The latter classification accuracy is directly related to determining the accuracy of Proficiency classifications for NCLB. These classification accuracy values were generally in the mid- to upper-eighties.

Table 12.8.1
Consistency Indexes for Performance Levels for the 2006 NJ ASK 2006

Test	Grade	Three Achievement Levels (PP, P, and AP)			Two Achievement Levels (Pass [P+AP], Fail [PP])		
		Cut-scores	Kappa	φ	Cut-score	Kappa	φ
LAL	5	16,30	0.38	76%	16	0.61	91%
	6	20,32	0.41	71%	20	0.57	84%
	7	21,34	0.40	73%	21	0.60	87%
Math	5	18,30	0.56	74%	18	0.63	89%
	6	17,31	0.63	78%	17	0.69	87%
	7	13,26	0.57	74%	13	0.61	82%

[28] Huynh, H. (1976). On the reliability of decisions in domain-referenced testing, *Journal of Educational Measurement*. 13:253-264.

PART 13: VALIDITY

13.1 Introduction

The *Standards for Educational and Psychological Testing* states, "Ultimately, the validity of an intended interpretation of test scores relies on all the available evidence relevant to the technical quality of a testing program. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees," (page 17).[29] While this section summarizes evidence supporting claims as to the validity of NJ ASK performance scores, many parts of this Technical Report provide appropriate evidence for validity. Some of this evidence is cross-referenced below for added convenience. Given the procedural and empirical evidence available and the rationale presented below, valid performance standards-based interpretations and uses of the scores are generally supported.

The following begins with a review of important federal statutes requiring the NJ ASK for grades 5–7 and goes on to explain the purposes and intended uses of performance test scores, suggesting the value implications of performance scores for schools, teachers, students, and parents. Content-related evidence supporting validity is presented in terms of the adequacy and appropriateness of the state content standards and the representation of the content standards on the tests. Then, validity evidence based on the internal structure of NJ ASK is provided through a correlational analysis of NJ ASK content clusters with each other. Reference to specific Standards within the *Standards for Educational and Psychological Testing* are provided where appropriate.

13.2 Federal Authority for School Accountability

NJ ASK performance standards scores for grades 5-7 reflect student achievement with respect to performance standards Partially Proficient, Proficient, and Advanced Proficient. The United States Department of Education bases accountability on school achievement of Adequate Yearly Progress (AYP) in reading/language arts and mathematics. AYP is set in terms of percentages of all students, and all student groups, scoring Proficient or above on the NJ ASK mathematics and LAL assessments.

13.3 Purpose and Intended Uses of Test Performance Scores[30]

The NJ ASK was developed for the following purposes and uses:

[29] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington: APA.

[30] Standard 1.2 – The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct that the test is intended to assess should be clearly described (page 17).

- NJ ASK tests scores provide an indication of student progress toward achieving the knowledge and skills identified in the New Jersey's Core Curriculum Content Standards (CCCS) and the tests fulfill the requirements under the No Child Left Behind (NCLB) Act;
- Annual school-improvement planning for instructional programs and other efforts directed toward the increased achievement of students in grade-, school-, or district-level groups; school improvement to reduce performance-gaps between disaggregated minority/majority groups such as racial, gender, disability, poverty, and limited English proficiency;
- Student, teacher, and parent information concerning the academic performance levels of individual students. Such information should be considered with other information, including other assessments and teachers judgment.

The valid interpretation and appropriate use of NJ ASK scores is supported in a variety of ways, including the training and consultation provided by personnel of the New Jersey Department of Education and publications such as the *New Jersey Assessment of Skills & Knowledge Spring 2006 Test Coordinator Manual Grades 5–7* and this Technical Report. The training and documentation provided to test users helps them better administer, understand, and use test score results.

13.4 NJ ASK Test Scores

The NJ ASK for grades 5–7 are scaled in several ways: raw score points, Item Response Theory (IRT), and performance standard level (based on scale-score cuts). New Jersey actively promotes the use of performance level results, reporting them annually on each content test at the student, school, district and state levels. Individual student and average scale scores are also used, but should play a secondary role, generally interpreted with reference to their distance from performance-score cut points. Test results are reported for students as a whole as well as by student group including gender, ethnicity, disability, use of accommodations, English language proficiency, migrant status, and poverty. Scores are reported to schools and districts in the annually published reports (see Part 8: Reporting).

NJ ASK performance scores indicate that an individual student performs at the Partially Proficient, Proficient, and Advanced Proficient level in a content area. Performance standard descriptions associated with each level provide details of the performance that students have met or exceeded. No stakes for students or teachers are attached by the state to student-level scores. Teachers are counseled to interpret individual student scores only in the context of other assessment results and their own experience.

13.5 Content-Related Evidence of Validity[31]

Baker and Linn (2002)[32] suggest that "Two questions are central in the evaluation of content aspects of validity. Is the definition of the content domain to be assessed adequate and appropriate? Does the test provide an adequate representation of the content domain the test is intended to measure?" (p. 6). The following two sections help answer these two very important questions and also address Standard 1.6 of the *Standards for Educational and Psychological Testing*.

13.6 Appropriateness of Content Definition

In 1996, the New Jersey State Board of Education adopted the New Jersey Core Curriculum Content Standards, an ambitious framework for educational reform in the State's public schools. New Jersey's standards were created to improve student achievement by clearly defining what all students should know and be able to do at the end of thirteen years of public education. Since the adoption of those standards, the NJDOE has continuously engaged in discussion with educators, business representatives, and national experts about the impact of the standards on classroom practices. To assist teachers and curriculum specialists in aligning curriculum with the standards, the NJDOE provided local school districts with a curriculum framework for each content area. The frameworks provided classroom teachers and curriculum specialists with sample teaching strategies, adaptations, and background information relevant to each of the content areas.

The State Board wisely required that the standards be reviewed and revised every five years. The review process, begun in May 2001, involved teachers, school administrators, students, parents, and representatives from business, higher education, and the community. In addition, several content areas were reviewed by Achieve, Inc., and the Council of Chief State School Officers (CCSSO). In response to this unprecedented review, the 2004 New Jersey Core Curriculum Content Standards provide the level of specificity and depth of content that will better prepare students for post secondary education and employment. The standards are based on the latest research in each of the content areas and identify the essential core of learning for all students.

Since the adoption of the original 1996 New Jersey Core Curriculum Content Standards (CCCS), the New Jersey State Board of Education approved administrative code that implements all aspects of standards-based reform. N.J.A.C. 6A:8 requires districts to align all curriculum to the standards; ensure that teachers provide instruction according to the standards; ensure student performance is assessed in each content area; and provide teachers with opportunities for professional development that focuses on the standards.

---

[31] Standard 1.6 – When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified (page 18).

[32] Baker, E. L., & Linn, R. L. (2002). Validity Issues for Accountability Systems. Center for the Study of Evaluation. Technical Report 585, Los Angeles, CA.

13.7 Adequacy of Content Representation

Adequacy of the content representation of the NJ ASK is critically important because the tests must provide an indication of student progress toward achieving the knowledge and skills identified in the CCCS, and the tests must fulfill the requirements under NCLB.

Adequate representation of the content domains defined in the CCCS is assured through use of a test blueprint and a responsible test construction process. New Jersey performance standards, as well as the CCCS, are taken into consideration in the writing of multiple-choice and open-ended items and open-ended rubric development. Each test must align with and proportionally represent the sub domains of the test blueprint. Evidence to support the above was given in Part 2, Test Development Process, and Part 4, Item Analysis. Tables 2.3.1 through 2.3.12 in Part 2 provide a comparison of target test construction maps to actual test maps for LAL and mathematics. Inspection of these tables confirms that the target number of items for each sub domain was achieved. As a representative example of this match, Table 2.3.1 is repeated below, with the addition of the actual counts as found in Table 2.3.2 (in parentheses).

Table 2.3.1
Test Construction Map for Grade 5 Language Arts Literacy NJ ASK

Text types/Strand	Reading Selections	MC (Number of Items)	OE (Number of Items)	WT (Number of Items)	Total Points
Picture Prompt		0 (0)	0 (0)	1 (1)	5 (5)
Narrative	1 (1)				
AT		4-6 (6)	0-2 (1)	0 (0)	8-12 (10)
WT		4-6 (4)	0-2 (1)	0 (0)	8-12 (8)
Everyday Text	1 (1)				
AT		2-6 (2)	0-2 (2)	0 (0)	8-12 (10)
WT		4-8 (8)	0-2 (0)	0 (0)	8-12 (8)
Total Items		**20 (20)**	**4 (4)**	**1 (1)**	
Total Points		**20 (20)**	**16 (16)**	**5 (5)**	**41 (41)**

The contractor strives to equitably represent the CCCS on each test by balancing sub-domain coverage on each test, by proportionally representing items corresponding to Partially Proficient, Proficient, and Advanced Proficient performance categories on each test, and by matching item format to the requirements of the content and standards descriptions.

13.8 Validity Evidence Based on the Internal Structure of NJ ASK[33]

Because the NJ ASK testing program assesses student performance in several content areas using a variety of testing methods, it is important to study the pattern of relationships among the content areas and testing methods. One method for studying patterns of relationships to provide evidence supporting the inferences made from test scores is the multi-trait, multi-method matrix. Tables 13.8.1, 13.8.2, and 13.8.3 summarize Pearson correlation coefficients among test content domains and clusters by grade level. The correlations between clusters within a content area were generally found to be higher than the correlations between clusters across the two content areas.

Table 13.8.1
Grade 5 - Correlation Coefficients among Content Domains and Clusters

	Writing	Reading	LAL1	LAL 2	Math	Math 1	Math 2	Math 3	Math 4	Math 5
Writing	1.00									
Reading	0.49	1.00								
Interpreting (LAL1)	0.43	0.94	1.00							
Analyzing (LAL2)	0.49	0.92	0.73	1.00						
Math	0.45	0.72	0.68	0.67	1.00					
Number (Math 1)	0.39	0.63	0.59	0.59	0.85	1.00				
Geometry (Math 2)	0.31	0.51	0.47	0.47	0.73	0.52	1.00			
Patterns (Math 3)	0.37	0.60	0.56	0.55	0.83	0.62	0.50	1.00		
Data Analysis (Math 4)	0.39	0.63	0.59	0.59	0.86	0.63	0.52	0.61	1.00	
Prob. Solving (Math 5)	0.44	0.72	0.67	0.67	1.00	0.85	0.71	0.84	0.86	1.00

[33] Standard 1.11 – If the rationale for a test use or interpretation depends on premises about the relationships among parts of the test, evidence concerning the internal structure of the test should be provided.

Standard 1.12 – When interpretation of subscores, score differences, or profiles is suggested, the rationale and relative evidence in support of such interpretation should be provided. Where composite scores are developed, the basis and rationale for arriving at the composites should be given.

Table 13.8.2
Grade 6 - Correlation Coefficients among Content Domains and Clusters

	Writing	Reading	LAL1	LAL 2	Math	Math 1	Math 2	Math 3	Math 4	Math 5
Writing	1.00									
Reading	0.54	1.00								
Interpreting (LAL1)	0.50	0.91	1.00							
Analyzing (LAL2)	0.50	0.92	0.68	1.00						
Math	0.50	0.73	0.69	0.64	1.00					
Number (Math 1)	0.42	0.61	0.58	0.54	0.82	1.00				
Geometry (Math 2)	0.44	0.63	0.60	0.56	0.86	0.63	1.00			
Patterns (Math 3)	0.42	0.61	0.58	0.54	0.86	0.62	0.65	1.00		
Data Analysis (Math 4)	0.45	0.64	0.61	0.56	0.87	0.64	0.67	0.66	1.00	
Prob. Solving (Math 5)	0.50	0.72	0.69	0.64	1.00	0.82	0.86	0.87	0.87	1.00

Table 13.8.3
Grade 7 - Correlation Coefficients among Content Domains and Clusters

	Writing	Reading	LAL1	LAL 2	Math	Math 1	Math 2	Math 3	Math 4	Math 5
Writing	1.00									
Reading	0.54	1.00								
Interpreting (LAL1)	0.44	0.88	1.00							
Analyzing (LAL2)	0.53	0.94	0.67	1.00						
Math	0.44	0.64	0.57	0.60	1.00					
Number (Math 1)	0.37	0.52	0.46	0.49	0.86	1.00				
Geometry (Math 2)	0.35	0.50	0.44	0.47	0.81	0.59	1.00			
Patterns (Math 3)	0.38	0.55	0.50	0.52	0.87	0.64	0.61	1.00		
Data Analysis (Math 4)	0.40	0.57	0.51	0.53	0.83	0.60	0.56	0.64	1.00	
Prob. Solving (Math 5)	0.44	0.63	0.56	0.59	1.00	0.86	0.80	0.87	0.82	1.00

13.9 Additional Evidence for Validity of NJ ASK

Validity evidence related to other Standards is listed below:

Standard 1.5[34]

- The composition of the sample of examinees from which validity evidence was obtained is described in detail in Part 11, including major relevant sociodemographic characteristics. This information is imbedded within the Tables of Part 11. These tables also provide descriptive statistics for number correct raw score and for scale scores. Statistics include N-counts, means, standard deviations, minimum and maximum values, and a variety of data disaggregations, including student demographic group and District Factor Group (DFG).

Standard 1.7[35]

- Standard setting procedures, including the selection process and the characteristics of judges, is described in detail in Part 3.
- The 2006 NJ ASK open-ended questions and writing responses at grades 5, 6, and 7 required hand scoring by Measurement Incorporated (MI) personnel. The processes of selecting and training scorers, reading and scoring papers, and monitoring scoring are described in detail in Part 6.

Standard 1.13[36]

- The conditions under which the data were collected are described in Part 5. Information about the administration of NJ ASK is available in the *New Jersey Assessment of Skills & Knowledge Spring 2006 Test Coordinator Manual Grades 5–7*.

[34] Standard 1.5 - The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics.

[35] Standard 1.7 – When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

[36] Standard 1.13 - When validity evidence includes statistical analyses of test results, either alone or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance.

APPENDIX 2-1
CHECKLIST FOR FORMS DEVELOPMENT

Table 2-1
Checklist for Forms Development

Item Data
Target Sum Rasch is between 495 and 500. If Target is not met, manager is made aware.
As many items as possible have a p-value above 0.35 and below 0.90
As many items as possible have a pt. bis above 0.20
No item was used as a sample item.

Item Pool
No passages are from the operational NJPASS
No writing prompts are from the Operational NJPASS (with the possible exception of the breach form writing prompt).
The NJPASS FT linking items are not used across grades.

Item Distribution
Item standards are distributed equally throughout the test
There are a variety of indicators assessed in each standard
MC items are generally in passage order, and OE items are at the end of the passage sets. WT items are in the appropriate places.
Answer key distribution is nearly equal between answer choices: A B C D
There are NOT more than 2 MC items in a row with the same answer.

Name, Gender, and Ethnicity Distributions
Check gender distribution (number of passages or prompts which have a male and/or female): Male Female Both
Check ethnicity distribution (number of passages or prompts): Caucasian_Hispanic Asian_African American Other_____
There are NOT two or more items in the same session that have similar contexts.
There are NOT two or more items with similar answers or answer choices.
Sample items and test items do NOT clue each other.
Items do NOT have any fairness or sensitivity related to the names and contexts of the items.

APPENDIX 3-1
DEMOGRAPHIC BACKGROUND OF PANELISTS PARTICIPATING IN STANDARD SETTING

Table 3-1
Demographic Background of PLD Panelists by Content Area/Grade Level

		LAL 5	LAL6	LAL 7	Math 5	Math 6	Math 7
Gender	M	2	0	1	2	2	1
	F	7	9	9	7	8	9
Ethnicity	W	9	8	8	9	7	7
	AA		1	2		2	2
Years Exp.	1-4	0	2	1	0	1	0
	5-10	4	1	2	3	1	1
	11-14	1	1	3	3	1	2
	15-19	2	0	0	2	4	1
	20-24	2	1	0	1	3	3
	25-30	0	4	1	0	0	1
	30+	0	0	3	1	0	1
DFG	A	1	3	1	1	2	4
	B	1	0	1	2	1	0
	CD	1	1	0	1	1	2
	DE	1	1	3	1	2	0
	FG	1	0	2	1	1	2
	GH	2	2	2	1	2	1
	I	2	2	0	3	1	1
	other	0	0	1	0	0	0
Region	N	3	4	6	6	3	6
	S	6	5	4	4	7	4
	Urban	2	5	1	3	2	5
Position	Teacher	5	6	6	7	7	5
	Adm.	4	3	4	3	3	5
Degree	B.A.	1	3	3	6	2	5
	M.A.	8	6	7	4	8	3
	Ph.D.	0	0	0	0	0	1

Table 3-2
Demographic Background of Phase 1 Panelists by Content Area and Grade Level

		Language Arts Literacy				Mathematics				Overall	
Grade		5	6	7		5	6	7		N	%
Gender	M	3	3	1		4	4	4		19	16
	F	16	16	19		16	16	16		99	84
Ethnicity	W	16	18	17		14	13	12		90	76
	AA	2	1	2		2	6	4		17	14
	H/L	1	0	1		2	0	3		7	6
Years Exp.	1-4	2	0	2		1	2	2		9	8
	5-10	6	4	6		6	3	6		31	26
	11-14	4	4	4		5	4	6		27	23
	15-19	1	4	1		0	3	2		11	9
	20-24	2	2	4		1	2	0		11	9
	25-30	3	3	3		2	4	2		17	14
	30+	1	2	0		0	2	2		7	6
DFG	A	1	1	3		3	2	3		13	11
	B	3	1	3		2	2	1		12	10
	CD	3	4	2		2	3	2		16	14
	DE	3	6	1		2	2	3		17	14
	FG	1	0	5		4	4	4		18	15
	GH	3	1	1		2	4	3		14	12
	I	4	3	2		3	2	2		16	14
	J	0	2	1		1	1	1		6	5
	CS	1	1	2		1	0	1		6	5
Region	N	9	8	13		8	10	12		60	51
	S	10	11	7		12	10	8		58	49
Position	Teacher	18	15	17		14	17	15		96	81
	Adm.	1	4	3		6	3	5		22	19
Degree	B.A.	5	4	5		10	9	6		39	33
	M.A.	14	15	15		9	11	13		77	65
	Ph.D.	0	0	0		1	0	1		2	2

APPENDIX 3-2
PERFORMANCE LEVEL DESCRIPTORS

New Jersey Assessment of Skills and Knowledge (NJ ASK)
Performance Level Descriptors
Language Arts Literacy
Grade 5

Proficient

The student performing at the proficient level demonstrates an ability to analyze and critique text, synthesizing details in order to build understanding. The student constructs meaning recognizing literary elements and figurative language. The student recognizes author's purpose and applies prior knowledge to draw conclusions.

As a proficient writer, the student develops a central focus and organizes and connects ideas with relevant details. The student exhibits some variety in word choice and sentence structure, attempts narrative techniques, and gives some evidence of transitions while incorporating basic writing mechanics.

Advanced Proficient

As a reader, the fifth grade student is able to synthesize details in order to make connections and generate new ideas. The student utilizes literary elements and author's purpose to analyze text.

As an advanced proficient writer, the student establishes and maintains a strong central focus and elaborates supporting details to convey ideas effectively. The student will include narrative techniques using fluid transitions, strong appropriate word choice and sentence variety to purposefully engage the reader.

New Jersey Assessment of Skills and Knowledge (NJ ASK)
Performance Level Descriptors
Language Arts Literacy
Grade 6

Proficient

The student performing at the proficient level demonstrates an understanding of the literal and inferential aspects of the text. Students activate prior knowledge and sustain comprehension by questioning, clarifying and predicting. Students at this level extrapolate relevant details to determine main idea, infer word meaning, in context, and identify the author's purpose. The proficient student recognizes organizational patterns and textual conventions.

As a proficient writer, the student develops a single focus and supporting details within an organizational structure. The student writes for a variety of purposes keeping audience in mind. Students at this level provide support for opinions and conclusions using textual and literary elements effectively.

Advanced Proficient

As a reader, the sixth grade advanced proficient student extends meaning by making conscious connections to text, author, self, and others. The student synthesizes information and draws insightful conclusions. Students at this level demonstrate an understanding of the central theme.

As an advanced proficient writer, in addition to consistently demonstrating the qualities outlined for a proficient student, the advanced proficient student develops a logical progression of ideas in a fluent and cohesive voice. Students at this level take appropriate compositional risks.

New Jersey Assessment of Skills and Knowledge (NJ ASK)
Performance Level Descriptors
Language Arts Literacy
Grade 7

Proficient

Seventh-grade students performing at the proficient level are able to interact with a variety of texts. Proficient students demonstrate an understanding of literal and inferential levels, while making relevant use of text to support or explain self-generated text. Students are able to identify central themes, supporting details and organizational structures of text. Students monitor their understanding of text and identify purpose for reading. Students can interpret textual conventions and extrapolate and synthesize information from text.

Seventh-grade students proficient in writing are able to develop a central theme, supporting details, and an organizational structure. Students at this level establish a purpose for writing and provide support for opinions and conclusions. Students demonstrate control of textual and literary elements.

Advanced Proficient

As an advanced proficient reader, in addition to consistently demonstrating the qualities outlined for a proficient student, the advanced student identifies abstract themes while analyzing and evaluating text. The advanced student utilizes prior knowledge to extend aspects of the text.

As an advanced proficient writer, in addition to consistently developing a central theme, supporting details and organizational structure, the advanced student demonstrates sophisticated use of textual and literary elements. The advanced student effectively uses compositional risks.

New Jersey Assessment of Skills and Knowledge (NJ ASK)
Performance Level Descriptors
Mathematics
Grade 5

Proficient

The student performing at the proficient level demonstrates the ability to recognize, understand, and apply basic mathematical concepts, skills, and terminology to theoretical and real world situations. The student will be able to perform basic computational procedures, apply geometric properties and spatial relationships, interpret data and graphs, apply the concepts and methods of discrete mathematics, and use basic algebraic concepts and processes. The student will also be able to infer, reason, and estimate while problem solving. The student will demonstrate flexibility in selecting a successful process or strategy. The student will demonstrate basic understanding of mathematical concepts through written expression.

Advanced Proficient

The student performing at the advanced proficient level will consistently demonstrate the qualities outlined for proficient performance. In addition, the student will analyze methods for appropriateness, synthesize processes, and evaluate mathematical relationships. The student will demonstrate conceptual understanding by consistently providing clear and complete explanations. The student abstracts concepts for use in other applications and successfully forms conjectures.

New Jersey Assessment of Skills and Knowledge (NJ ASK)
Performance Level Descriptors
Mathematics
Grade 6

Proficient

The student performing at the proficient level in mathematics will be able to demonstrate evidence of and communicate conceptual understanding of procedural and analytical skills. The student applies mathematical skills and knowledge to theoretical and real world situations. In addition, the student makes connections across mathematical domains. The student at this level understands and applies appropriate standard numerical operations and estimations—an understanding sufficient for problem solving in practical situations. The student will be able to determine the reasonableness of an answer. The student understands and applies basic geometric concepts including properties, measurement, and spatial relationships. The student clearly interprets data and graphs, determines probabilities, applies the concepts and methods of discrete mathematics, and uses basic algebraic concepts and processes.

Advanced Proficient

The student performing at the advanced proficient level in mathematics will consistently demonstrate the qualities for proficient performance. In addition, the student at the advanced level demonstrates the use of abstract thinking and provides explanations that are consistently clear and thorough. The student will support a logical efficient method to solving the problem. The student consistently makes accurate inferences and predictions. The student supports responses by using appropriate mathematical terminology. The student successfully analyzes and draws appropriate inferences from data.

New Jersey Assessment of Skills and Knowledge (NJ ASK)
Performance Level Descriptors
Mathematics
Grade 7

Proficient

A proficient student demonstrates a conceptual understanding of mathematical knowledge, procedures, skills, and processes across the four content standards. A student will demonstrate an ability to:

- Identify, recognize, and compare different representations of numbers and numerical operations.
- Identify, describe, classify, apply, and solve problems involving geometry, spatial sense, and measurement.
- Recognize, identify, and extend simple patterns and algebraic representations of theoretical and real-world problems.
- Model situations, solve problems, analyze, and draw appropriate inferences from data.
- Understand fundamental concepts of probability and discrete mathematics.

Advanced Proficient

An advanced proficient student consistently demonstrates the qualities outlined for proficient performance. In addition, the student consistently demonstrates the ability to:

- Think abstractly using inductive and deductive reasoning.
- Use a variety of problem solving strategies.
- Extrapolate information; form and support conclusions through clear and thorough explanation(s).
- Assess reasonableness of the solution(s).

APPENDIX 7-1
RAW SCORE TO SCALE SCORE TABLES

Table 7-1
Grade 5 Regular Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
5	0	100
5	1	100
5	2	100
5	3	102
5	4	114
5	5	124
5	6	133
5	7	140
5	8	147
5	9	154
5	10	160
5	11	166
5	12	171
5	13	176
5	14	181
5	15	186
5	16	191
5	17	196
5	18	200
5	19	204
5	20	209
5	21	213
5	22	217
5	23	221
5	24	225
5	25	229
5	26	232
5	27	237
5	28	241
5	29	245
5	30	250
5	31	255
5	32	261
5	33	267
5	34	275
5	35	283
5	36	295
5	37	300
5	38	300
5	39	300

Table 7-2
Grade 6 Regular Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
6	0	100
6	1	100
6	2	118
6	3	131
6	4	140
6	5	148
6	6	154
6	7	160
6	8	165
6	9	170
6	10	174
6	11	178
6	12	182
6	13	186
6	14	190
6	15	193
6	16	197
6	17	200
6	18	203
6	19	207
6	20	210
6	21	213
6	22	216
6	23	220
6	24	223
6	25	227
6	26	230
6	27	234
6	28	238
6	29	242
6	30	246
6	31	250
6	32	255
6	33	260
6	34	265
6	35	272
6	36	280
6	37	291
6	38	300
6	39	300

Table 7-3
Grade 7 Regular Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
7	0	100
7	1	100
7	2	122
7	3	136
7	4	147
7	5	155
7	6	163
7	7	169
7	8	175
7	9	181
7	10	186
7	11	191
7	12	196
7	13	200
7	14	204
7	15	208
7	16	213
7	17	217
7	18	220
7	19	224
7	20	228
7	21	232
7	22	235
7	23	239
7	24	243
7	25	246
7	26	250
7	27	254
7	28	258
7	29	262
7	30	267
7	31	272
7	32	278
7	33	285
7	34	292
7	35	300
7	36	300
7	37	300
7	38	300
7	39	300

Table 7-4
Grade 5 Braille Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
5	0	100
5	1	100
5	2	100
5	3	107
5	4	119
5	5	129
5	6	138
5	7	146
5	8	153
5	9	159
5	10	165
5	11	171
5	12	177
5	13	182
5	14	187
5	15	192
5	16	200
5	17	201
5	18	206
5	19	210
5	20	214
5	21	218
5	22	222
5	23	226
5	24	231
5	25	235
5	26	239
5	27	243
5	28	250
5	29	253
5	30	259
5	31	265
5	32	273
5	33	282
5	34	293
5	35	300
5	36	300
5	37	300

Table 7-5
Grade 6 Braille Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
6	0	100
6	1	100
6	2	120
6	3	132
6	4	142
6	5	149
6	6	156
6	7	162
6	8	167
6	9	172
6	10	176
6	11	180
6	12	184
6	13	188
6	14	192
6	15	195
6	16	200
6	17	202
6	18	205
6	19	209
6	20	212
6	21	215
6	22	219
6	23	222
6	24	226
6	25	229
6	26	233
6	27	237
6	28	241
6	29	245
6	30	250
6	31	254
6	32	259
6	33	264
6	34	271
6	35	279
6	36	290
6	37	300
6	38	300

Table 7-6
Grade 7 Braille Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
7	0	100
7	1	100
7	2	124
7	3	138
7	4	149
7	5	158
7	6	165
7	7	172
7	8	178
7	9	184
7	10	189
7	11	194
7	12	200
7	13	203
7	14	208
7	15	212
7	16	216
7	17	221
7	18	225
7	19	229
7	20	234
7	21	238
7	22	243
7	23	250
7	24	252
7	25	257
7	26	263
7	27	268
7	28	275
7	29	282
7	30	289
7	31	298
7	32	300
7	33	300
7	34	300
7	35	300

Table 7-7
Grade 5 Large Print Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
5	0	100
5	1	100
5	2	100
5	3	107
5	4	119
5	5	129
5	6	138
5	7	146
5	8	153
5	9	159
5	10	165
5	11	171
5	12	177
5	13	182
5	14	187
5	15	192
5	16	200
5	17	201
5	18	206
5	19	210
5	20	214
5	21	218
5	22	222
5	23	226
5	24	231
5	25	235
5	26	239
5	27	243
5	28	250
5	29	253
5	30	259
5	31	265
5	32	273
5	33	282
5	34	293
5	35	300
5	36	300
5	37	300

Table 7-8
Grade 6 Large Print Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
6	0	100
6	1	100
6	2	120
6	3	132
6	4	142
6	5	149
6	6	156
6	7	162
6	8	167
6	9	172
6	10	176
6	11	180
6	12	184
6	13	188
6	14	192
6	15	195
6	16	200
6	17	202
6	18	205
6	19	209
6	20	212
6	21	215
6	22	219
6	23	222
6	24	226
6	25	229
6	26	233
6	27	237
6	28	241
6	29	245
6	30	250
6	31	254
6	32	259
6	33	264
6	34	271
6	35	279
6	36	290
6	37	300
6	38	300

Table 7-9
Grade 7 Large Print Mathematics Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
7	0	100
7	1	100
7	2	124
7	3	138
7	4	149
7	5	158
7	6	165
7	7	172
7	8	178
7	9	184
7	10	189
7	11	194
7	12	200
7	13	203
7	14	208
7	15	212
7	16	216
7	17	221
7	18	225
7	19	229
7	20	234
7	21	238
7	22	243
7	23	250
7	24	252
7	25	257
7	26	263
7	27	268
7	28	275
7	29	282
7	30	289
7	31	298
7	32	300
7	33	300
7	34	300
7	35	300

Table 7-10
Grade 5 Regular LAL Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
5	0	110
5	1	132
5	2	146
5	3	154
5	4	161
5	5	166
5	6	170
5	7	174
5	8	178
5	9	181
5	10	184
5	11	187
5	12	190
5	13	193
5	14	195
5	15	198
5	16	200
5	17	202
5	18	205
5	19	207
5	20	210
5	21	213
5	22	216
5	23	219
5	24	222
5	25	226
5	26	230
5	27	234
5	28	239
5	29	244
5	30	250
5	31	257
5	32	265
5	33	275
5	34	287
5	35	300
5	36	300
5	37	300
5	38	300
5	39	300
5	40	300
5	41	300

Table 7-11
Grade 6 Regular LAL Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
6	0	100
6	1	100
6	2	100
6	3	108
6	4	119
6	5	127
6	6	135
6	7	142
6	8	148
6	9	153
6	10	159
6	11	164
6	12	169
6	13	173
6	14	177
6	15	182
6	16	186
6	17	189
6	18	193
6	19	197
6	20	200
6	21	203
6	22	207
6	23	210
6	24	214
6	25	217
6	26	221
6	27	225
6	28	229
6	29	234
6	30	239
6	31	244
6	32	250
6	33	256
6	34	263
6	35	271
6	36	279
6	37	288
6	38	298
6	39	300
6	40	300
6	41	300
6	42	300
6	43	300
6	44	300
6	45	300
6	46	300
6	47	300
6	48	300

Table 7-12
Grade 7 Regular LAL Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
7	0	100
7	1	100
7	2	108
7	3	121
7	4	130
7	5	137
7	6	144
7	7	150
7	8	155
7	9	159
7	10	164
7	11	168
7	12	172
7	13	176
7	14	179
7	15	182
7	16	186
7	17	189
7	18	192
7	19	194
7	20	197
7	21	200
7	22	203
7	23	206
7	24	209
7	25	212
7	26	215
7	27	218
7	28	222
7	29	225
7	30	229
7	31	234
7	32	239
7	33	244
7	34	250
7	35	257
7	36	264
7	37	272
7	38	281
7	39	291
7	40	300
7	41	300
7	42	300
7	43	300
7	44	300
7	45	300
7	46	300
7	47	300
7	48	300

Table 7-13
Grade 5 Braille LAL Raw Score to Scale Score Table

Grade	Raw Score	Scale Score
5	0	110
5	1	133
5	2	146
5	3	155
5	4	161
5	5	166
5	6	171
5	7	175
5	8	179
5	9	182
5	10	186
5	11	189
5	12	192
5	13	195
5	14	200
5	15	201
5	16	204
5	17	208
5	18	211
5	19	214
5	20	218
5	21	221
5	22	225
5	23	230
5	24	235
5	25	240
5	26	250
5	27	254
5	28	264
5	29	277
5	30	294
5	31	300
5	32	300
5	33	300
5	34	300
5	35	300
5	36	300

Table 7-14
Grade 6 Braille LAL Raw Score to Scale Score Table

SAME AS TABLE 7-11

Table 7-15
Grade 7 Braille LAL Raw Score to Scale Score Table

SAME AS TABLE 7-12

Table 7-16
Grade 5 Large Print LAL Raw Score to Scale Score Table

SAME AS TABLE 7-13

Table 7-17
Grade 6 Large Print LAL Raw Score to Scale Score Table

SAME AS TABLE 7-11

Table 7-18
Grade 7 Large Print LAL Raw Score to Scale Score Table

SAME AS TABLE 7-12